UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA E ENGENHARIA DE ALIMENTOS

DISTRIBUIÇÃO DE TAXAS DE RECOMBINAÇÃO AO LONGO DO CROMOSSOMO 4 DE *Arabidopsis thaliana* E SUA ASSOCIAÇÃO COM ELEMENTOS GENÔMICOS

ADILSON SANTOS MARTINS

Orientador:
Prof. Alexandre Siqueira Guedes Coelho

Goiânia - GO
Brasil

Março - 2010

ADILSON SANTOS MARTINS

DISTRIBUIÇÃO DE TAXAS DE RECOMBINAÇÃO AO LONGO DO CROMOSSOMO 4 DE *Arabidopsis thaliana* E SUA ASSOCIAÇÃO COM ELEMENTOS GENÔMICOS

Tese apresentada ao Programa de Pós-graduação em Agronomia, da Universidade Federal de Goiás, como requisito parcial à obtenção do título de Doutor em Agronomia, área de concentração: Genética e Melhoramento de Plantas.

Orientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Goiânia, GO - Brasil
2010

Dados Internacionais de Catalogação na Publicação (CIP)
GPT/BC/UFG

Martins, Adilson Santos.
M386d Distribuição de taxas de recombinação ao longo do cromossomo 4 de *Arabidopsis thaliana* e sua associação com elementos genômicos [manuscrito] / Adilson Santos Martins. – 2010.
138 f. : il.

Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho.
Tese (Doutorado) – Universidade Federal de Goiás, Escola de Agronomia e Engenharia de Alimentos, 2010.
Bibliografia: f. 124-133.

1. *Arabidopsis thaliana* – Melhoramento Genético 2.Plantas – Melhoramento Genético I. Título.

CDU: 575.111:635.3

Termo de Ciência e de Autorização para Disponibilizar as Teses e Dissertações Eletrônicas (TEDE) na Biblioteca Digital da UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás–UFG a disponibilizar gratuitamente através da Biblioteca Digital de Teses e Dissertações – BDTD/UFG, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: [ ] Dissertação [X] Tese

2. Identificação da Tese ou Dissertação:

Autor(a):	**ADILSON SANTOS MARTINS**				
CPF:	342.378.791-00	E-mail:	adilson@pucgoias.edu.br		
Seu e-mail pode ser disponibilizado na página?	[X]Sim [ ] Não				
Vínculo Empregatício do(a) Autor(a):	Pontifícia Universidade Católica de Goiás – PUC Goiás				
Agência de fomento:				Sigla:	
País:	Brasil	UF:	GO	CNPJ:	01.587.609/0001-71
Título:	Distribuição de taxas de recombinação ao longo do cromossomo 4 de *Arabidopsis thaliana* e sua associação com elementos genômicos.				
Palavras-chave:	Permuta; melhoramento genético; DSB; desequilíbrio de ligação				
Título em outra língua:	Distribution of recombination rates across the chromosome 4 of *Arabidopsis thaliana* and its association with genomic features				
Palavras-chave em outra língua:	Crossover; DSB; linkage disequilibrium;				
Área de concentração:	Genética e melhoramento de plantas				
Data defesa:	29/março/2010				
Programa de Pós-Graduação:	Programa de Pós-Graduação em Agronomia da EA da UFG				
Orientador(a):	Prof. Dr. Alexandre Siqueira Guedes Coelho				
CPF:	491.567.801-68	E-mail:	coelho@agro.ufg.br		
Co-orientador(a):					
CPF:		E-mail:			
Co-orientador(a):					
CPF:		E-mail:			

3. Informações de acesso ao documento:

Liberação para disponibilização?[1] [X] total [ ] parcial

Em caso de disponibilização parcial, assinale as permissões:

[ ] Capítulos. Especifique: _____

[ ] Outras restrições: _____

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF não-criptográfico da tese ou dissertação.

O Sistema da Biblioteca Digital de Teses e Dissertações garante aos autores, que os arquivos contendo eletronicamente as teses e ou dissertações, antes de sua disponibilização, receberão procedimentos de segurança, criptografia (para não permitir cópia e extração de conteúdo, permitindo apenas impressão fraca) usando o padrão do Acrobat.

Adilson Santos Martins

Assinatura do(a) autor(a)

Data: 29/março/ 2010

[1] Em caso de restrição, esta poderá ser mantida por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Todo resumo e metadados ficarão sempre disponibilizados.

ADÍLSON SANTOS MARTINS

TÍTULO: **"Distribuição de taxas de recombinação ao longo do cromossomo 4 de *Arabidopsis thaliana* e sua associação com elementos genômicos"**

Tese DEFENDIDA e APROVADA em 29 de março de 2010, pela Banca Examinadora Constituída pelos membros:

Prof. João Batista Duarte
EA/UFG

Prof. Rinaldo Wellerson Pereira
Universidade Católica de Brasília

Dr. Flávio Breseghello
Embrapa Arroz e Feijão

Prof. Lázaro José Chaves
EA/UFG

Prof. Alexandre Siqueira Guedes Coelho
Presidente – EA/UFG

Goiânia - Goiás
Brasil

À minha querida esposa Cleide Sônia, à minha filha Débora, ao meu filho Artur e aos meus pais Deocleciano e Aldaires,

DEDICO.

AGRADECIMENTOS

A Deus, por ter me dado a oportunidade de conviver e aprender com todos essas pessoas, abaixo citadas ou não.

À Profª MSc. Maria Silvia Monteiro, minha primeira professora de genética;

Ao Prof. Dr. Lázaro José Chaves, pessoa que se tornou minha referência como docente e como lidar com as relações interpessoais;

Ao Prof. Dr. Alexandre Siqueira Guedes Coelho, pelos incontáveis momentos oportunizados para debates enriquecedores, por me acolher como seu orientado e pela dedicação e disponibilidade com a qual sempre me atendeu;

À Profª Dra. Patrícia Guimarães Santos Melo, digníssima coordenadora do PPGA, pela atenção e paciência com a qual sempre nos atendeu.

Aos membros da banca de defesa, Prof. Rinaldo Wellerson Pereira, DSc (UCB), Flávio Breseghello, PhD (Embrapa/CNPAF), Rosana Pereira Vianello Brondani, PhD (Embrapa/CNPAF), Prof. Dr. Lázaro José Chaves (EA/UFG), Prof. Dr. João Batista Duarte (EA/UFG) e o Prof. Dr. Américo José dos Santos Reis (EA/UFG), nosso sincero agradecimento pelo precioso tempo dedicado para colaborar conosco.

Aos pesquisadores Dr. Damian Labuda, do Departamento de Pediatria da Universidade de Montreal, e ao bioinformata Jean-François Lefebvre, do Centro de Pesquisas do Hospital *Sainte-Justine*, em Montreal, no Canadá, pelas discussões construtivas de ajustes do *software* InfRec utilizado para cálculos das estimativas da recombinação e, também, pela disponibilização de hora/CPU do servidor Unix daquela instituição.

Aos pesquisadores Dr. Jason A. Greenbaum, do Programa de Bioinformática, e Dr. Thomas D. Tullius, do Departamento de Química, ambos da Universidade de Boston – EUA, pela gentileza de terem processado, nos computadores da Universidade de Boston, as sequências dos pseudocromossomos de *A. thaliana*, calculando as estimativas da intensidade de clivagem por radical OH^- em cada nucleotídeo.

Ao secretário do PPGA, Welinton Barbosa Mota, pela atenção dedicada, pela presteza em seus serviços e, principalmente, pela amizade e boa companhia em vários momentos;

Aos meus irmãos, minhas cunhadas e minha sogra, pelos constantes estímulos durante esta caminhada.

SUMÁRIO

RESUMO

MARTINS, A. S. **Distribuição de taxas de recombinação ao longo do cromossomo 4 de *Arabidopsis thaliana* e sua associação com elementos genômicos**. 2010. 138 f. Tese (Doutorado em Agronomia: Genética e Melhoramento de Plantas) – Escola de Agronomia e Engenharia de Alimentos, Universidade Federal de Goiás, Goiânia, 2010[1].

A recombinação é um processo chave na evolução da organização dos genomas das espécies, importante para garantir a segregação adequada dos cromossomos homólogos durante a meiose I e criar novas combinações de alelos, gerando variabilidade genética para a ação da seleção natural. Do ponto de vista molecular, a recombinação é iniciada por uma lesão na fita dupla de DNA, denominada *Double-Strand Break* (DSB), seguida da formação de uma junção dupla de Holliday (dHJ), a qual é resolvida por vias alternativas. Quando há troca de material genético entre os cromossomos homólogos caracteriza-se a ocorrência de um evento de recombinação, *crossover*, visualizado citogeneticamente por meio de um quiasma. Estudos citológicos, genéticos e moleculares realizados em vários organismos demonstraram que a distribuição de *crossover* ao longo dos cromossomos não é regular, mas concentrada em fragmentos relativamente pequenos de DNA, denominados *hotspots* de recombinação. Na busca por correlações entre a distribuição de elementos genômicos e a de ocorrência de *hotspots* um modelo ajustado com dados do genoma humano se mostrou capaz de explicar até 42% da variação na taxa de recombinação, numa escala de 5 mega pares de bases. Em plantas, apesar da existência de vários genomas já sequenciados nenhum trabalho nesse sentido ainda foi realizado, pelo menos na ordem de resolução proporcionada pela recente disponibilidade de dados genéticos obtidos com o uso de *chips* de alta densidade de marcas SNP. Usando dados genéticos de populações, obtidos por genotipagem de 362 acessos de *A. thaliana* com 250 mil marcas SNP, estimativas da intensidade de clivagem por radical OH^- e dados da sequência de nucleotídeos do cromossomo 4 de *A. thaliana* o presente trabalho propõe-se a: *i*) caracterizar a distribuição de taxas de recombinação e de desequilíbrio de ligação ao longo do cromossomo 4, em várias escalas; *ii*) identificar fragmentos *hotspots* de recombinação; e *iii*) identificar elementos genômicos com provável associação à ocorrência desses *hotspots*. Os resultados obtidos mostraram que a distribuição das taxas de recombinação ao longo do cromossomo 4 de *A. thaliana* é bastante concentrada, pois proporções entre 50% e 60% dos eventos de recombinação ocorrem em apenas 13% a 20% da sequência de DNA. Variáveis genômicas como a porcentagem da soma das bases G e C (G+C%) e a intensidade de clivagem por radical OH^- apresentam correlações significativas com as estimativas do desequilíbrio de ligação em várias escalas. A média da intensidade de clivagem por radical OH^- proporciona informação redundante com a variável G+C%. O uso de janelas deslizantes sobrepostas gera distroções que provocam o surgimento artificial de correlações fortes e significativas. Os fragmentos *hotspots* de recombinação têm uma distribuição concentrada no terço médio do cromossomo, enquanto os fragmentos caracterizados por longo alcance do desequilíbrio de ligação estão localizados, predominantemente, nos terços distais.

Palavras-chave: permuta, *crossover*, desequilíbrio de ligação, genoma, DSB.

ABSTRACT

MARTINS, A. S. **Distribution of recombination rates across the chromosome 4 of *Arabidopsis thaliana* and its association with genomic features**. 2010. 138 f. Thesis (Doctor in Agronomy: Genetics and Plant Breeding) - Escola de Agronomia e Engenharia de Alimentos, Universidade Federal de Goiás, Goiânia, 2010[1].

Recombination is one of the most important factors in the evolution of genome organization. It provides the links between homologous chromosomes that ensure their proper segregation during the first meiotic division. It is responsible for the creation of novel allele combinations and yields genetic diversity on which evolutionary selection can act. Double-strand DNA breaks (DSB) initiate meiotic recombination and when the 3' terminus of one of the broken strands invades the unbroken DNA molecule and primes DNA synthesis a double Holliday junction must be resolved through some alternative pathways. When homologous chromosomes exchange genetic material with each other, an event of recombination or a crossover takes place, which may be seen through chiasma. Citological, genetics, and molecular studies in many organisms have demonstrated that crossovers have a non homogeneous distribution across chromosomes, and rather concentrated in relative small DNA fragments usually called recombination hotspots. In searching for genomic features associated with recombination hotspots a model fitted to human genome data explained 42% of recombination rate variation in a 5 mega base pairs scale. Despite the fact that genomes of some plant species have been already sequenced, up to this moment, no research has been published concerning a high resolution characterization of recombination rate variation across a plant's genome. This study used OH^- radical cleavage intensity estimates and sequence data of chromosome 4 of *A. thaliana* and population genetic data from a public set of 250 thousand SNP genotypes obtained for 362 *A. thaliana* accessions to: *i*) characterize the recombination rate and linkage disequilibrium (LD) distributions across the chromosome 4 in different scales; *ii*) search for recombination hotspots; *iii*) evaluate probable associations between sequence motifs and genomic features with recombination hotspots. The results have shown that the distribution of recombination events across chromosome 4 of *A. thaliana* is very concentrated: 50% to 60% of all recombination events spans in only 13% to 20% of the total length of the chromosome. Genomic features as G+C percent (G+C%) and OH^- radical cleavage intensity showed important associations with LD estimates in several scales. The mean OH^- radical cleavage intensity and G+C% showed redundancy in correlation analysis with LD and recombination rates. Artificial strong and statistically significant correlations arose from the usage of sliding windows. DNA fragments considered as hotspots lay preferentially in the middle third of the chromosome, while those characterized for having long range LD decay are most localized in the two distal thirds of the chromosome.

Key words: recombination, linkage disequilibrium, genome, DSB, crossover.

1 INTRODUÇÃO

A recombinação é uma das forças evolutivas fundamentais que contribuem para a determinação do padrão de variação dos polimorfismos ao longo do genoma. A caracterização da variação da taxa de recombinação ao longo do genoma, em diferentes escalas, tem implicações diretas nos estudos de evolução, de mapeamento por associação e no entendimento das bases moleculares da recombinação. A inferência da variação da taxa de recombinação ao longo do genoma consiste num dos desafios mais importantes da moderna genética de populações (Wang & Rannala, 2008).

À medida que se amplia o conhecimento sobre o padrão de variação da taxa de recombinação ao longo do genoma, amplia-se também o entendimento dos padrões de associação entre alelos de locos diferentes, ou desequilíbrio de ligação (LD). Consequentemente, serão ampliadas as possibilidades de uso desse entendimento para o mapeamento das causas genéticas da variação fenotípica. A maior ou menor facilidade para identificar os componentes genéticos responsáveis pela variação fenotípica depende do grau de conhecimento que se tem sobre as associações existentes entre diferentes regiões do genoma, associações que são determinadas, predominantemente, pelo processo de recombinação (Stumpf & McVean, 2003).

Os métodos de determinação, ou medição, das taxas de recombinação em maiores níveis de resolução são complicados, demorados e dispendiosos. Os métodos de construção de mapas genéticos e de mapeamento de locos de caracteres quantitativos (*Quantitative Trait Loci* – QTL) usam delineamentos genéticos baseados em cruzamentos controlados (*ex.:* retro cruzamentos, populações F2, linhagens puras recombinantes, linhagens duplo-haplóides, cruzamentos entre heterozigotos). A densidade de marcas e a quantidade de indivíduos que, em geral, é utilizada nesses métodos não propiciam um nível de amostragem de meioses informativas compatível com a determinação da variação das taxas de recombinação em alta resolução, na escala de genes individuais ou de alguns milhares de pares de bases. Experimentos usando cruzamentos em larga escala podem ser proibitivamente dispendiosos.

Neste contexto, Stumpf & McVean (2003) sugerem o uso de métodos

estatísticos indiretos para acessar informações sobre taxas de recombinação, tais como os métodos que usam dados genéticos de populações. Esses métodos inferem taxas de recombinação a partir de padrões de variação genética entre sequências de DNA de indivíduos amostrados aleatoriamente de uma população.

Nessa linha, inúmeros métodos foram propostos, desenvolvidos e, em seguida, implementados em programas de computador (Fearnhead & Donnelly, 2001; Hudson, 2001; McVean et al., 2002; Fearnhead & Donnelly, 2002; Li & Stephens, 2003; McVean et al., 2004; Fearnhead & Smith, 2005; Fearnhead, 2006; Auton & McVean, 2007; Wang & Rannala, 2008; Lefebvre & Labuda, 2008). Os métodos baseados em funções de verossimilhança completa ou composta são computacionalmente muito intensivos. Com o aprimoramento destas metodologias e o concomitante aumento da capacidade de memória e de processamento dos computadores atuais, espera-se que isso deixe de ser uma limitação à aplicação desses métodos. A abordagem utilizada por Lefebvre & Labuda (2008), por exemplo, é bem menos exigente em esforço computacional e proporciona estimativas confiáveis da variação da taxa de recombinação.

Hayashi & Iwata (2010) comentam que os avanços recentes das tecnologias de sequenciamento e o desenvolvimento dos *microarrays* (coleção de pontos microscópicos de fragmentos de DNA depositados de maneira ordenada sobre uma superfície sólida e fixados a ela por meio de ligações covalentes) proporcionaram a descoberta e a análise de polimorfismos ao longo dos genomas de várias espécies, incluindo plantas, animais e outros organismos. Sistemas de genotipagem em grande escala, como os *microarrays* de SNP (*Single Nucleotide Polimorphisms*) de alta densidade, que permitem a genotipagem de centenas de milhares de marcas SNP distribuídas ao longo de todo genoma, têm sido utilizados de forma eficiente para se avaliar indivíduos de uma amostra a custos que a cada dia estão se tornando mais acessíveis. Antecipando este contexto, Meuwissen et al. (2001) propuseram uma nova tecnologia de melhoramento genético que utiliza dados de marcadores SNP em alta densidade, distribuídos em todo o genoma.

Para alguns organismos, a disponibilidade de *microarrays* de alta densidade para genotipagem SNP ao longo de todo o genoma já é uma realidade. No caso de *Arabidopsis thaliana*, Kim et al. (2007) genotiparam 19 acessos usando um *microarray* contendo 341.602 marcas SNP. Pelo padrão observado da variação do desequilíbrio de ligação ao longo do genoma, esses autores concluiram que um número de 250.000 marcas SNP seria adequado para a realização de estudos de associação em escala genômica, nesta

espécie. Estes *microarrays*, com 250.000 marcas SNP distribuídas ao longo do genoma, já vêm sendo comercializados pela empresa Affymetrix.

Os termos *hotspots* e *coldspots* de recombinação são comumente utilizados para descrever regiões de um genoma com "altas" e "baixas" ocorrências de eventos de recombinação, respectivamente. Porém, como comentam Hellenthal & Stephens (2006), tem sido difícil observar uma homogeneidade de critérios entre os vários métodos de quantificação e classificação das medidas associadas a esses termos. Os resultados oriundos de diferentes estudos são difíceis de serem comparados pelo fato de que as medidas associadas a esse termos são sensíveis às suposições do que vem a ser *hotspot* ou *coldspot*. DeMassy (2003), por exemplo, considera como sendo um *hotspot* de recombinação um fragmento que, apesar de apresentar taxa de recombinação igual à média do genoma, está localizada em uma região que apresenta taxas de recombinação relativamente baixas. Ao se tomar a média de uma região como referência local para classificar um segmento de DNA como *hotspot* pode resultar em enormes discordâncias entre resultados de diferentes autores.

Dooner (1986) usou o termo *hotspot* de recombinação para designar o loco do gene *bronze*, que apresentava uma taxa de recombinação 100 vezes maior que a média de todo o genoma de milho. Em revisão sistematizada por Lichten & Goldman (1995) o termo *hotspot* foi usado para indicar um loco ou uma região que apresenta uma taxa de recombinação maior do que a média observada no genoma e, *coldspot* para designar as regiões com taxas de recombinação menores do que a média do genoma. Petes (2001) elaborou uma revisão na qual é apresentada uma classificação dos *hotspots*: *α-hotspots* são os *hotspots* cuja ativação é dependente da ligação de fatores detranscrição específicos; *β-hotspots* são aqueles posicionados em regiões ricas em sequências que repelem a formação de nucleossomos e os *γ-hotspots* designa os *hotspots* que estão associados a regiões com picos na porcentagem da soma das bases G e C. Auton & McVean (2007), tomando como referência a taxa de recombinação média em uma determinada região, definem *hotspot* como variações na taxa de recombinação que se caracterizam por picos elevados em relação àquela média local. A inovação proposta por Auton & McVean (2007) é a caracterização quantitativa dos *hotspots*, estimando e atribuindo-lhes uma posição, uma extensão e um nível de intensidade. Ainda não há um consenso na literatura especializada sobre a definição do seja um *hotspot*, nem dos atributos que melhor poderiam o descrever.

Usando dados genéticos de populações humanas, McVean et al. (2004)

caracterizaram a variação da taxa de recombinação ao longo do genoma com resolução nunca antes atingida, destacando-se o padrão da distribuição dos eventos de recombinação e a intensidade e localização dos *hot spots* e *cold spots* de recombinação. Em seguida, Myers et al. (2005), usando dados genéticos obtidos da genotipagem de 1,6 milhões de marcas SNP, em 71 indivíduos de três populações, refizeram a caracterização das taxas de recombinação ao longo do genoma humano para identificar fragmentos *hotspots* e *coldspots* de recombinação. Dentro desses fragmentos foram realizadas buscas extensivas por elementos genômicos que mostrassem alguma associação com a variação da taxa de recombinação. Os elementos genômicos identificados e mais significativos foram utilizados num modelo linear para realizar a predição da taxa de recombinação. Na escala de 5 Mb, o modelo ajustado explicava 42% da variação na taxa de recombinação. Para as escalas de 500 kb, 50 kb e 5 kb os modelos explicaram 34%, 15% e 4%, respectivamente.

Desde a publicação do trabalho de Kim et al. (2007), vem sendo disponibilizado dados de genotipagem SNP em alta densidade, usando o chip de 250.000 marcas fabricado pela Affymetrix. Na versão publicada em dezembro de 2009 constavam a genotipagem de 362 acessos de *A. thaliana*. Alguns trabalhos publicados têm feito uso desse tipo de dados: Kim et al. (2007) caracterizaram o decaimento do desequilíbrio de ligação ao longo do genoma, usando dados de genotipagem de 19 acessos; e Nordborg et al. (2005) e Clark et al. (2007) caracterizaram a variação dos níveis de polimorfismo ao longo do genoma. Apesar da disponibilidade destes dados genéticos populacionais em alta densidade para *A. thaliana*, até o presente momento, nenhuma caracterização das taxas de recombinação e/ou do desequilíbrio de ligação ao longo do genoma, objetivando a identificação de *hotspots* e *coldspots* de recombinação foi publicada. Drouaud et al. (2006) buscaram por associações entre as taxas de recombinação e a distribuição de elementos genômicos usando a técnica de decomposição em componentes principais; mas, eles utilizaram dados oriundos da genotipagem de apenas 71 SNP ao longo do cromossomo 4.

Com base nos resultados obtidos por Myers et al. (2005), mostrando a diminuição do poder preditivo dos modelos com a diminuição da escala de observação dos dados, e nos resultados de Kendal & Suomela (2005), que evidenciam aumento no valor absoluto das correlações entre as distribuições dos elementos genômicos à medida em que há aumento da escala, a busca por associação entre a distribuição dos elementos genômicos e a variação na taxa de recombinação deve ser realizada considerando-se esta escala.

Usando dados da sequência de nucleotídeos do cromossomo 4 de *A. thaliana*,

estimativas da intensidade de clivagem por radical OH^- e, ainda, dados genéticos obtidos por genotipagem do cromossomo 4, usando-se 41761 marcadores SNP avaliados em 362 acessos de *A. thaliana*, o presente trabalho propôs-se a: *i*) caracterizar a distribuição de taxas de recombinação e de desequilíbrio de ligação ao longo do cromossomo 4, em várias escalas; *ii*) identificar fragmentos *hotspots* de recombinação; e *iii*) identificar elementos genômicos com provável associação à ocorrência desses *hotspots*.

2 REVISÃO DE LITERATURA

2.1 RECOMBINAÇÃO

2.1.1 Importância do processo de recombinação na evolução do genoma de plantas

Observa-se grande variação nos genomas nucleares das plantas, quanto à algumas características como o seu tamanho, o número de cromossomos, a densidade de genes e a ordem dos genes, por exemplo. Gaut et al. (2007) lembram que, no caso da família das gramíneas (Poaceae), a divergência teve início há aproximadamente 77 milhões de anos e resultou em diferenças de até 55 vezes nos tamanhos dos genomas diplóides e de até dez vezes no número diplóide de cromossomos. Eventos de poliploidia, deleção de genes, rearranjos genômicos, duplicação de genes, sem dúvida, contribuíram na moldagem do atual estado dos genomas das plantas.

A importância relativa de cada tipo de evento durante a evolução das plantas varia de grupo para grupo. Em *Arabidopsis thaliana*, por exemplo, a proporção de genes que foram duplicados por meio de eventos localizados de duplicação é comparável à proporção de duplicação provocada por eventos de poliploidia. A duplicação localizada de genes pode sofrer uma aceleração após um evento de poliploidia, mas também funcionam na sua ausência (Bennetzen et al., 2005).

Na opinião de Gaut et al. (2007), um processo que não tem recebido a atenção adequada pelos estudiosos da evolução das plantas é o processo de recombinação, que gera mutações e influencia a força da seleção natural. Esses autores destacam algumas evidências que apontam para um importante papel da recombinação na evolução do genoma das plantas. Entre essas evidências os autores destacam: as propriedades mutacionais da recombinação, o papel da recombinação na seleção natural e a relação entre a organização estrutural dos genomas e o padrão de variação da recombinação ao longo dos genomas.

A recombinação, em geral, é iniciada em um ponto de ocorrência de quebra da dupla fita de DNA, ou um *Double Strand Break*, referido como DSB de agora por diante,

seguido de resolução deste DSB com troca de fragmentos entre as cromátides homólogas. A resolução de um DSB pode gerar vários tipos de mutações, principalmente quando a recombinação acontece em sequências repetitivas não perfeitamente alinhadas. Como exemplo, Gaut et al. (2007) citam o caso em que resoluções desbalanceadas de DSB entre cromátides irmãs ou entre cromossomos homólogos podem gerar o aumento ou a diminuição do número de cópias de um elemento repetitivo, como ilustrado na Figura 1.

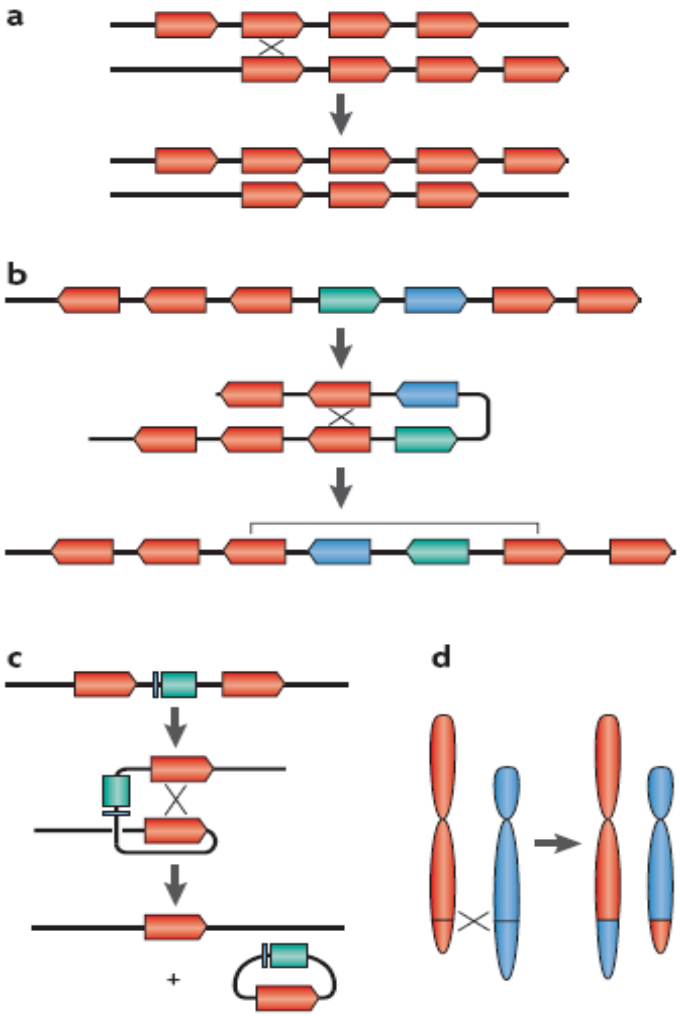


Figura 1. Tipos de mutações que podem ser geradas por recombinação em regiões de sequências repetidas. Os símbolos vermelhos representam repetições de mesma sequência. Em a, b e c cada linha representa uma fita de DNA de uma cromátide. **a**) uma recombinação não simétrica entre cromossomos homólogos (ou entre cromátides irmãs) em regiões de sequências repetitivas não alinhadas gera um aumento do número de repetições em um cromossomo e diminuição no número de repetições no outro cromossomo; **b**) uma recombinação intra-fita em uma região de repetições pode gerar uma inversão de um fragmento de DNA; **c**) uma recombinação intra-fita entre dois motivos repetitivos pode gerar uma deleção; **d**) quando ocorre uma recombinação homóloga entre regiões repetitivas de cromossomos diferentes pode gerar uma translocação. Fonte: Gaut et al. (2007).

A ocorrência de eventos de recombinação não simétrica é relativamente comum. Jelesko et al. (2004) apresentaram dados que mostram a ocorrência de novas variantes no número de cópias de sequências repetitivas entre cromátides irmãs a uma taxa aproximada de 10^{-6} por conjunto gênico, por planta F1 e por meiose, em pesquisa realizada num conjunto que agrupa três genes em *A. thaliana*. Considerando que esta espécie tem aproximadamente 1500 conjuntos de dois ou três genes e que a taxa de 10^{-6} por conjunto gênico é semelhante em todos os conjuntos, espera-se que pelo menos uma em cada 700 sementes possa conter uma variação no número de cópias de sequências repetitivas. E esta é uma medida conservadora, pois despreza todas as outras possíveis causas de geração de variação em número de cópias de sequências repetitivas. A conclusão de Gaut et al. (2007) é a de que o processo de recombinação por si só é responsável pela geração de um número substancial de variação no número de cópias de sequências repetitivas nas plantas.

Porém, mutações não são causadas somente por eventos de recombinação não simétrica entre cromossomos homólogos. Recombinações intra-cromátides, em regiões com elementos repetitivos posicionados em uma mesma direção ou posicionados de forma invertida, podem gerar inversões (Figura 1-b), deleções (Figura 1-c) e, talvez mais frequentemente, conversões gênicas. Quando um evento de recombinação ocorre entre regiões genômicas distintas, é gerada uma translocação de segmentos cromossômicos (Figura 1-d). Evidências que apontam a recombinação como fonte de ocorrência de translocações são apresentadas por Lysak et al. (2006), em trabalhos com espécies do gênero *Brassica* e em estudo realizado no nível de sequências de nucleotídeos, conduzido por Ziolkowski et al. (2003). Ambos mostram a ocorrência de translocações entre cromossomos de *A. thaliana*. Lockton & Gaut (2005) sugerem que as chances de ocorrência de translocações em plantas é bem maior do que em outros eucariotos devido ao fato de que as plantas têm famílias de genes maiores e mais numerosas, as quais consistem em regiões adequadas para a ocorrência desse tipo de fenômeno. Nesse sentido, Gaut et al. (2007) ressaltam que o papel da recombinação na geração de polimorfismos pontuais em nucleotídeos ainda é um assunto que merece mais estudos.

Uma vez que a recombinação tenha gerado mutações e variabilidade, sua ação continua desempenhando papel importante por meio de seu efeito na seleção natural. Para as regiões genômicas com alta taxa de recombinação, as mutações vantajosas rapidamente aumentam sua frequência populacional e as deletérias são eliminadas. Para ilustrar esse caso, os autores Gaut et al. (2007) apresentam a seguinte situação exemplo: suponha que

duas mutações benéficas surjam em dois locos distintos de um mesmo homólogo de um cromossomo, uma delas num indivíduo, e a outra num segundo indivíduo da mesma população; ocorrendo recombinação ambas as mutações poderão ser fixadas simultaneamente pela ação da seleção natural; não ocorrendo recombinação, as duas mutações não poderão se juntar num mesmo indivíduo e, neste caso, ocorrerá uma competição entre as duas mutações e elas não terão chances de serem fixadas de forma simultânea, diminuindo a eficiência da seleção natural.

As taxas de recombinação variam entre regiões genômicas, tanto em animais quanto em plantas. Porém, são poucas as espécies para as quais os padrões de variação da taxa de recombinação ao longo dos genomas já são conhecidos e estão disponíveis. Para as espécies que se tem estimativas da taxa de recombinação ao longo do genoma observa-se uma característica geral: a ocorrência da diminuição ou supressão quase total da taxa de recombinação nas regiões heterocromáticas. A proporção da heterocromatina no genoma e o grau de supressão da recombinação varia de espécie para espécie (Gaut et al., 2007). Por exemplo, Kim et al. (2005) informaram que as regiões heterocromáticas recobrem 50% do genoma de sorgo e nestas regiões as taxas de recombinação são, em média, 34 vezes menores do que nas regiões de eucromatina. Trabalhos conduzidos por Wang et al. (2006) e Tanksley et al. (1992) sobre o genoma do tomateiro caracterizaram que 75% deste genoma está em regiões de heterocromatina pericentromérica, nas quais ocorre uma supressão da ordem de mil vezes na taxa de recombinação, quando se compara com as taxas das regiões eucromáticas. Os mecanismos que governam a supressão heterocromática das taxas de recombinação ainda não são totalmente elucidados, porém, podem estar relacionados com modificações de natureza epigenética, conforme informam Yan et al. (2005). Esse padrão de variação da taxa de recombinação em escala ampla pode ser bastante heterogêneo em escalas menores. Gaut et al. (2007) citam o trabalho de Fu et al. (2002), no qual se verificaram taxas de recombinação dentro do gene *bronze* em milho, que podem ser até cem vezes maiores do que a média geral no genoma, e que a diferença das taxas de recombinação entre a parte distal e proximal do gene difere em duas ordens de grandeza. Isso significa que os níveis das taxas de recombinação variam não apenas em escala cromossômica, mas também em escala de kilo pares de bases (kb).

Nesse contexto, Gaut et al. (2007) lançaram a pergunta: será que os diferentes tipos de elementos genômicos estão organizados de acordo com gradientes de taxas de recombinação ao longo dos genomas das plantas? Para responder a esta pergunta esses

autores recorreram a algumas informações como: *i*) a ocorrência de maior densidade de elementos transponíveis (TE) e menor densidade de genes em regiões de baixas taxas de recombinação; *ii*) a ocorrência de duplicações locais associadas a regiões de altas taxas de recombinação; e *iii*) o efeito da recombinação na evolução da ordem de genes.

O acúmulo relativamente maior de elementos genéticos móveis (elementos transponíveis, ou simplesmente TE, de *Transposable Elements*) e de DNA oriundo de organelas em regiões heterocromáticas com baixas taxas de recombinação - conforme pesquisas conduzidas por Matsuo et al. (2005) e por Bowers et al. (2005) - indica que essas elementos genômicos são lentamente eliminados pela seleção. Gaut et al. (2007) ressaltam que, atualmente, ainda não se tem uma descrição completa das forças que governam a acumulação dos TE; mas, há razões para se considerar que os padrões de acumulação dos TE são influenciados por pressões de seleção que variam com a taxa de recombinação. Dados, sobre o genoma do arroz (*Oryza sativa*), apresentados pelo projeto "*International Rice Genome Sequencing Project*", IRGSP (2005), explicitam que os TE que se acumulam em regiões pobres em genes e com baixas taxas de recombinação tendem a ser maiores que os TE acumulados em regiões ricas em genes e com altas taxas de recombinação (Figura 2).

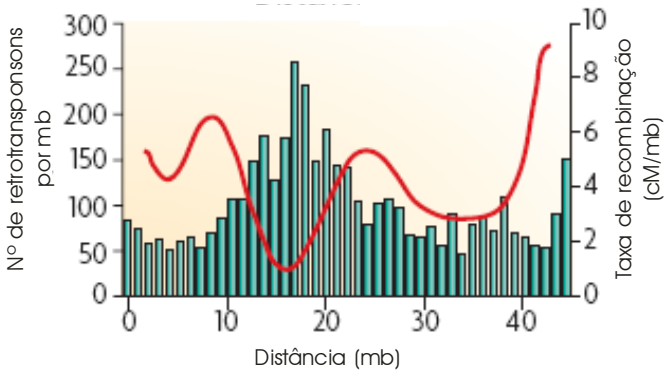


Figura 2. Correlação entre a quantidade de elementos transponíveis e a taxa de recombinação ao longo do cromossomo 1 de arroz (*Oriza sativa*). Os elementos genômicos estão totalizados em janelas de 100 kb. A linha vermelha representa a variação da taxa de recombinação, suavizada. Fonte: IRGSP (2005).

Os TE de maiores tamanhos ao se transporem para regiões de altas taxas de recombinação e ricas em genes têm maiores chances de provocar rupturas de genes ou funções genômicas e, por isso, serem eliminados pela seleção natural com maior eficiência. Os TE menores gozariam de uma probabilidade menor de provocar tais rupturas. Quando

TE maiores se transpõem para regiões de menor densidade de genes, com baixas taxas de recombinação, as chances de provocar rupturas no funcionamento genômico são menores e, por isso, eles se acumulariam nessas regiões em maior proporção (Gaut et al., 2007).

Por outro lado, se esses elementos não funcionais não se acumularem em regiões de altas taxas de recombinação, então essas regiões se enriqueceriam com elementos genômicos funcionais. Vários resultados que corroboram esta idéia foram publicados: dados do IRGSP (2005), sobre o genoma do arroz (*Oriza sativa*); Wright et al. (2003), sobre o genoma de *A. thaliana*; Anderson et al. (2006), sobre o genoma do milho (*Zea mays*); Dvorak et al. (2004) sobre o genoma do trigo (*Triticum aestivum*); Wang et al. (2006) sobre o genoma do tomateiro (*Solanum lycopersicum*). No caso dos trabalhos com o milho e tomate, os resultados mostraram que a densidade de genes é de quatro a dez vezes menor nas regiões heterocromáticas do que nas de eucromatina. Especificamente para espécies como o arroz e o milho, a correlação entre taxa de recombinação e densidade de genes é mantida quando se consideram as análises dentro de regiões eucromáticas.

Ao ser considerado que as duplicações gênicas locais podem ser originadas a partir de eventos de recombinação não simétricos, os genes duplicados deveriam se agrupar em regiões com altas taxas de recombinação. E este é o caso dos conjuntos de genes em tandem, que são bem mais abundantes em regiões com altas taxas de recombinação em *A. thaliana* (Figura 3), como mostrado por Zhang & Gaut (2003). Em arroz, isto é mostrado por Rizzon et al. (2006) e, em trigo, por Akhunov et al. (2003), mesmo após a correção para a variação na densidade de genes. De acordo com Thomas (2006), os conjuntos de genes em tandem estão distribuídos de forma muito similar no organismo *Caenorhabditis elegans*, indicando que a maior densidade de grupos de genes em tandem em regiões com altas taxas de recombinação pode ser uma característica generalizável entre os eucariotos. A distribuição dos conjuntos de genes em tandem pode ser resultado da interação entre a ocorrência dos eventos de recombinação não simétrica, em regiões com altas taxas de recombinação, com os mecanismos que governam as taxas de perdas e retenções dos genes dentro desses conjuntos.

Dvorak & Akhunov (2005), em estudo com trigo e espécies aparentadas, verificaram que tanto os eventos de duplicação, quanto os de deleção, ocorrem de maneira mais frequente em regiões de altas taxas de recombinação. As taxas de duplicação são até três vezes maiores nas regiões distais, com altas taxas de recombinação, do que nas regiões proximais que possui taxas de recombinação mais baixas. Gaut et al. (2007) argumentam

que o fato de este padrão ser observado tanto em uma espécie diplóide quanto numa poliplóide reforça a idéia de que a seleção purificadora não é tão efetiva em altos níveis de poliploidia e corrobora a idéia de que a recombinação tem um papel importante na ocorrência dos fenômenos de duplicação e de deleção.

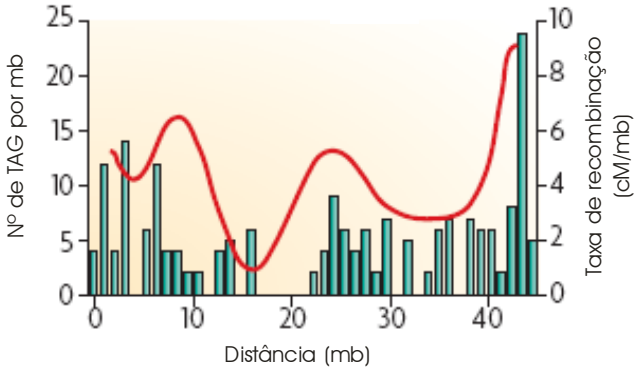


Figura 3. Correlação entre a quantidade de grupos de genes em tandem (TAG do inglês *Tandem Arrayed Genes*) e a taxa de recombinação ao longo do cromossomo 1 de *A. thaliana*. Os elementos genômicos estão totalizados em janelas de 100 kb. A linha vermelha representa a variação da taxa de recombinação, suavizada. Fonte: Zhang & Gaut (2003).

Além de sua atuação na distribuição e localização das deleções, duplicações e de outros elementos estruturais e funcionais dos genomas, o processo de recombinação pode ser considerado como responsável pela evolução da ordem dos genes ou, em outras palavras, do grau de sintenia entre espécies com algum ancestral em comum. No estudo de Dvorak & Akhunov (2005), foi verificado que as regiões teloméricas e outras regiões eucromáticas com altas taxas de recombinação sofrem evolução mais rápida na ordem dos genes, devido, em grande parte, à alta incidência de duplicações e deleções. Diante dessas informações Gaut et al. (2007) sugeriram que a ordem dos genes pode ser melhor conservada em regiões de média a baixa taxas de recombinação. Nestas regiões as taxas de mutação também são relativamente baixas, embora as taxas de recombinação estão em níveis adequados para permitir uma ação mais eficiente da seleção. Enfim, sumarizam Gaut et al. (2007), a taxa de evolução da ordem dos genes varia com a espécie, com a historia das populações e com a localização genômica, mas tudo como função das taxas de recombinação. Eles reforçam sua posição a favor de que a recombinação é uma força evolutiva que molda alguns aspectos da variabilidade nos genomas de plantas.

2.1.2 As diferentes rotas da recombinação homóloga

A recombinação homóloga promove variabilidade genética e contribui para que ocorra uma segregação correta dos cromossomos durante a meiose. Porém, a garantia de segregação adequada dos cromossomos homólogos requer um conjunto de eventos de alta complexidade. Parte dessa complexidade envolve justamente a estimulação da recombinação homóloga durante a meiose I (Whitby, 2005). A importância da recombinação, na maioria dos organismos, implica em, pelo menos, duas notáveis funções: i) promover maior diversidade genética por meio da geração de novas combinações entre os alelos; ii) estabelecer as conexões físicas (quiasmas) entre os homólogos, as quais são imprescindíveis para a ligação bipolar aos fusos durante a meiose I.

Como resultado de estudos em diversos organismos, predominantemente em fungos, alguns modelos têm sido propostos para os desdobramentos dos eventos da recombinação homóloga. Um dos mais importantes é o modelo de reparo das quebras na dupla fita de DNA (*Double-Strand Break repair model*), ou DSB, proposto originalmente por Szostak et al. (1983) e ilustrado na Figura 4. De acordo com esse modelo a recombinação é iniciada pela formação de um DSB em um dos homólogos (o homólogo vermelho e magenta na Figura 4, passo 1). As extremidades 5' das fitas de DNA que foram cortadas sofrem degradação de um pequeno fragmento gerando duas caldas de fitas simples com terminações 3' (Figura 4, passo 2). Uma dessas caldas (a de cor magenta) encontra sua sequência complementar em uma das fitas do cromossomo homólogo (a de cor azul), conforme ilustrado na Figura 4, passo 3; e penetra entre as duas por meio de uma bolha de denaturação denominada de D-loop (*Displacement loop*). Outra denominação a essa penetração de uma extremidade 3' na bolha de denaturação é invasão de fita simples, ou SEI, do inglês *Single End Invasion*. A fita invasora serve de iniciadora para a síntese de DNA que se extende até encontrar a outra extremidade 3' do outro lado do DSB (Figura 4, passos 4 e 5). Ao capturar a segunda extremidade 3', as quatro fitas aneladas formam o que se denomina dupla junção de Holliday (*Double Holliday Junction*), daqui por diante simplesmente denominada de dHJ (Figura 4, passo 6).

Quando a resolução de uma dHJ se dá por meio da quebra do mesmo par de fitas em cada ponto de junção (posições a><a e b><b) ocorre um evento não-*crossover* e não há formação de quiasma. Por outro lado, quando a resolução da dHJ acontece com a quebra de fitas diferentes em cada ponto de junção (por exemplo, posições a><a e B), cada

um dos cromossomos homólogos conterá DNA próprio de um lado da junção e DNA do seu homólogo do outro lado, caracterizando a ocorrência de *crossover* (Figura 4, passo 7).

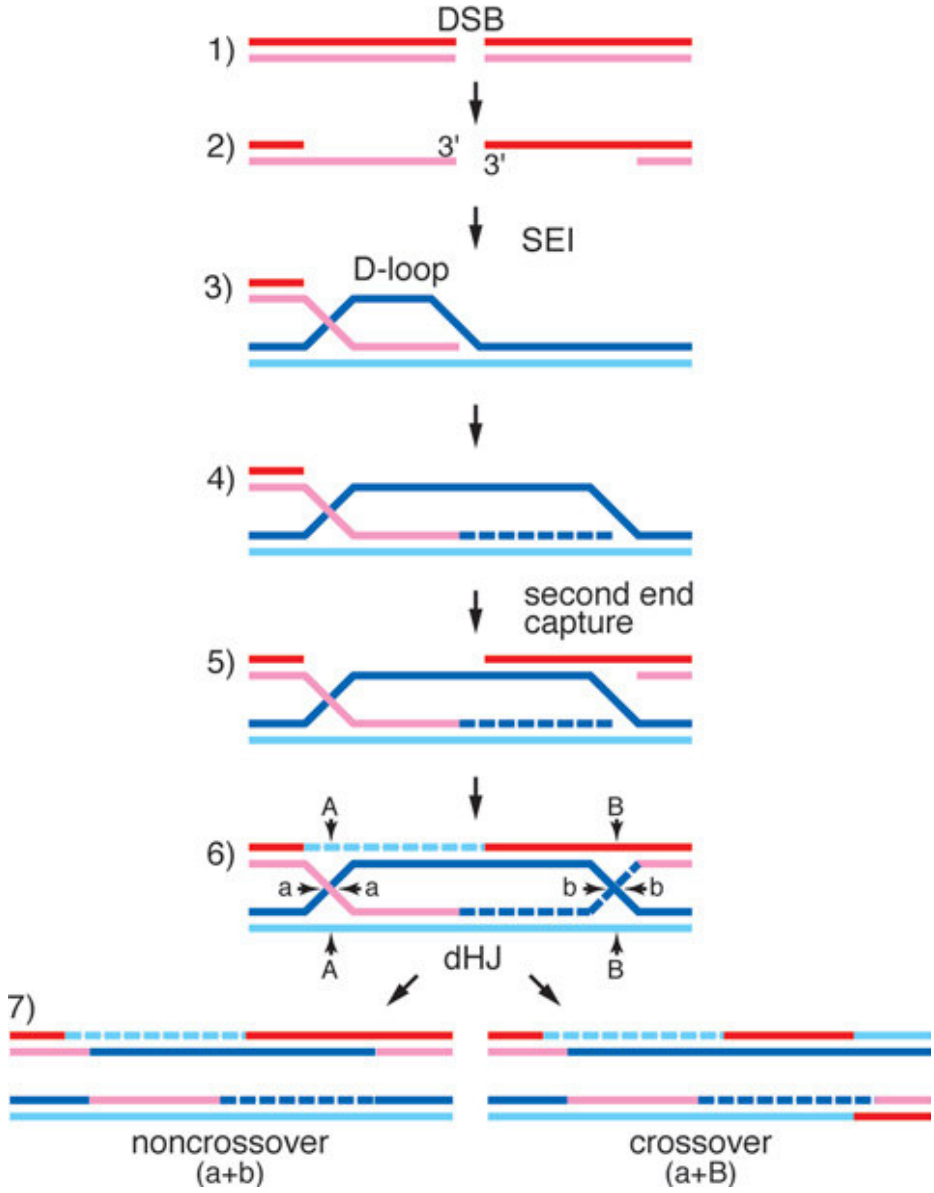


Figura 4. Modelo de Szostak et al. (1983) para o reparo das quebras na dupla fita de DNA (DSB). *SEI*: invasão da bolha de denaturação pela extremidade 3' da fita simples de DNA. *D-loop*: bolha de DNA dupla fita denaturado. *Second end capture*: anelamento da segunda extremidade 3' de fita simples e subseqüente síntese de DNA. *dHJ*: dupla junção de Holliday. Linhas tracejadas representam DNA recém sintetizado. Eventos de recombinação (*crossover*) podem ocorrer com quebras de fitas diferentes nas posições A+b ou a+B. Quebras das mesmas fitas nas posições a+b ou A+B não geram eventos de recombinação (*noncrossover*). Fonte: Whitby (2005).

Desde que o modelo de reparo de DSB foi proposto por Szostak et al. (1983), várias enzimas que catalizam reações que ocorrem nos vários estágios do reparo de DSB foram identificadas. Verificou-se que a grande maioria dessas proteínas compartilha uma

mesma história evolutiva, apresentando alto grau de conservação (Lin et al., 2006). Dentre elas incluem-se a Spo11, que gera os DSB, o complexo Mre11-Rad50-Xrs2, que é responsável pela degradação das extremidades 5' das fitas simples, e a Rad51, que, juntamente com Dmc1 e outras proteínas acessórias, comanda as reações de pareamento dos homólogos e de invasão da fita simples, para gerar as dHJ (Whitby, 2005).

As nucleases que quebram as dHJ não estão, entretanto, completamente identificadas e caracterizadas. Além disso, Bishop & Zickler (2004) verificaram que em *Sacharomices cerevisiae* algumas mutações provocam a redução da resolução de dHJ e da formação de *crossover*, sem afetar a formação de DSB e a geração de eventos não-*crossover*. Isto dá indícios de que os *crossover* e os não-*crossover* podem se derivar de rotas independentes de reparo dos DSB. Existem alternativas de modelos que propõem que as dHJ sejam resolvidas, exclusivamente, na via que gera *crossover* (Figura 5, passos 6a e 7a), enquanto os eventos não-*crossover* seriam gerados por um mecanismo denominado anelamento de fita dependente de síntese, ou SDSA (*Synthesis-Dependent Strand Annealing*), pelo qual a a dissociação da extremidade invasora após a síntese de DNA é sucedida pelo anelamento das duas extremidades não invasoras com mais síntese de DNA (Figura 5, passos 3b a 6b).

A definição de qual dos caminhos serão tomados para o reparo de DSB depende do conjunto de proteínas específicas da meiose que atuarão sobre determinada posição onde está localizado um DSB. Certo conjunto dessas proteínas, denominado geralmente de ZMM (Zip1, Zip2, Zip3, Mer3, Msh4 e Msh5), parece promover a estabilidade da terminação 3' invasora e a formação de dHJ, de forma a conduzir o reparo do DSB pela via que gera *crossover*. A planta *A. thaliana* contém essas proteínas e, portanto, uma das vias de resolução de DSB conduz à geração de *crossover*. A distribuição desses *crossover* não é totalmente aleatória, havendo tendência de estes não se posicionarem próximos uns aos outros. A via dependente do complexo ZMM é tida como sinônima da via que gera *crossover* com interferência (Bishop & Zickler, 2004).

Muller[1] (1916), citado por Mézard et al. (2007), definiu interferência: "A ocorrência de um *crossover* interfere na ocorrência de outro *crossover* num mesmo par de cromossomos e eu denominei esse fenômeno de 'interferência'". Uma das conseqüências da manifestação da interferência é que os *crossover* tendem a se posicionarem mais

[1] Muller, H. J. The mechanisms of crossing-over. **American Naturalist**, n. 592, v. 50, p. 193-221, abr., 1916.

distantes uns dos outros do que seria esperado se o seu posicionamento fosse independente. Uma explicação para a ocorrência da interferência seria o fato de que as subsequentes ligações dessas enzimas às posições dos *crossover* geraria uma diminuição na concentração local, diminuindo sua ligação a pontos vizinhos ao local do *crossover*. Outra explicação seria a de que a ligação das enzimas a um determinado local enviaria um sinal para a vizinhaça que inibiria a resolução de DSB próximo pela via dependente do complexo ZMM. Esse sinal poderia ser a imposição e liberação do estresse aplicado aos eixos dos homólogos.

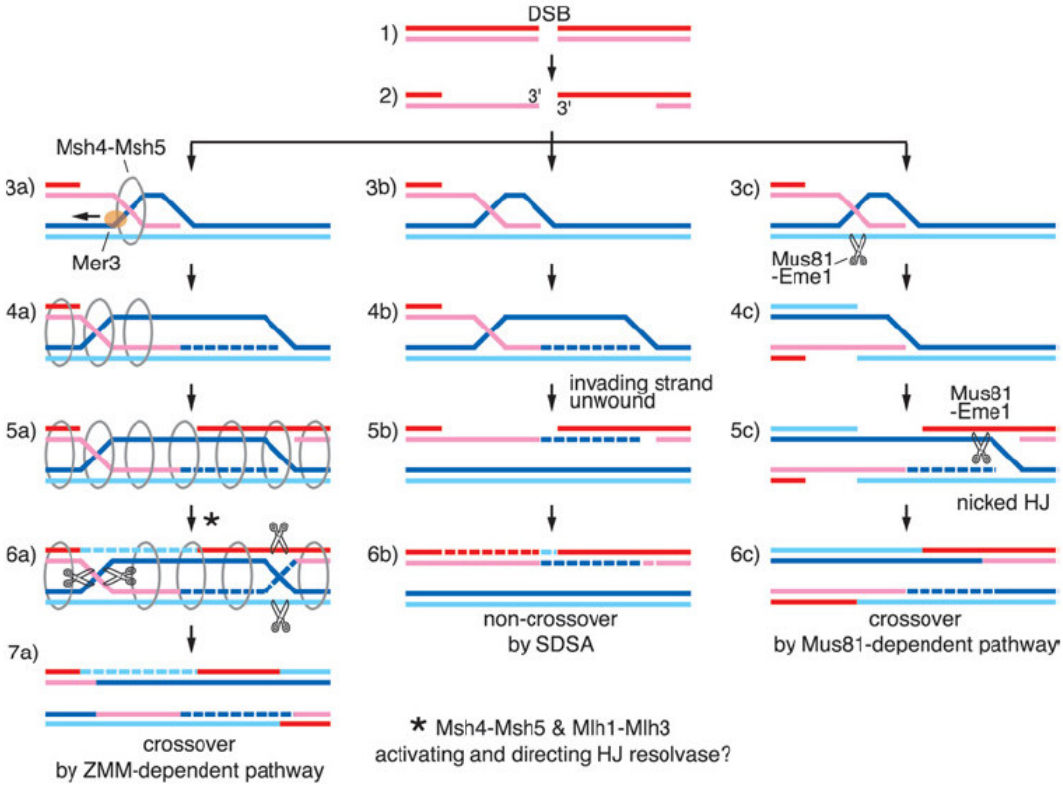


Figura 5. Vias de resolução das quebras na dupla fita de DNA (DSB). *ZMM*: grupo de proteínas (Zip1, Zip2, Zip3, Mer3, Msh4 e Msh5) específicas da meiose envolvidas no reparo de DSB e de *dHJ*. *dHJ*: dupla junção de Holliday. A coluna à esquerda representa a via de resolução de DSB e de dHJ que depende do complexo ZMM para gerar eventos *crossover* e, portanto, sujeito aos efeitos da interferência. A coluna à direira representa a via que depende do complexo Mus81-Eme1 para gerar eventos *crossover* e, portanto, não sofre o efeito da interferência. A coluna ao centro representa a via que depende do complexo Msh4-Msh5 e Mlh1-Mlh3 e gera eventos não-*crossover*. *Invading strand unwound*: a fita simples invasora, após o anelamento e a síntese de DNA com o molde do homólogo se liberta e propicia a troca não recíproca de material genético. *SDSA*: anelamento de fita simples dependente de síntese de DNA. Linhas tracejadas representam DNA recém sintetizado. Fonte: Whitby (2005).

Nem todos os *crossover* são, entretanto, resultantes da via dependente do

complexo ZMM. Existe uma via que é comandada pelas proteínas Mus81 e Mms4. Em alguns organismos como no *Schizosacharomices pombe*, a formação de *crossover* é totalmente derivada da via que dependente da proteína Mus81 (Smith et al., 2003). Nesse organismo verifica-se a ausência do complexo ZMM e de *crossover* com interferência.

Na ausência ou sem a ação das proteínas ZMM, as extremidades 3' se tornariam instáveis e transientes. Bishop & Zickler (2004) sugerem que estas estruturas instáveis são canalizadas para vias alternativas, como a do SDSA.

Na maioria dos organismos a via ZMM ou outra equivalente, que geram *crossover* com interferência, parece ser a principal via utilizada para a formação dos *crossover*. Esta parece ser a única via empregada pelos mamíferos. Em alguns organismos, entre os quais está *A. thaliana*, além da via ZMM, também é utilizada outra via, que depende de Mus81, em pelo menos uma parte do total dos *crossover*. Mercier et al. (2005) e Mézard et al. (2007) confirmaram a coexistência das duas vias em *A. thaliana*.

Para resumir, em um extremo poderiam ser agrupados organismos como o nematóide *C. elegans*, nos quais a totalidade dos eventos de *crossover* ocorre sob efeito de interferência, comandada pela via dependente do complexo ZMM. No outro extremo, em que se inclui o fungo *S. pombe*, estariam os organismos cujo posicionamento de *crossover* não está sujeito aos efeitos da interferência, e que são regulados pela via das proteínas Mus81 e Mms4. Entre os dois extremos estaria o grupo de organismos que faz uso de ambas as vias de regulação da recombinação ou, talvez, outras vias também. Em fungos, tomateiro e *A. thaliana*, a maioria dos *crossover* ocorre pela via ZMM e sob efeito da interferência, mas parte deles escapa da interferência pela via dependente de Mus81 e Mms4. Por exemplo, em fungos e tomateiro cerca de 30% dos *crossover* ocorrem sem interferência; já em *A. thaliana* estima-se que 15% deles escapam da via com interferência (Mézard et al., 2007).

2.1.3 Distribuição dos eventos de recombinação ao longo dos cromossomos

A ocorrência de um *crossover* envolve a permuta de grandes fragmentos de material genético entre cromossomos homólogos durante a meiose. A quantidade e a distribuição dos *crossover* ao longo dos cromossomos estão sob forte controle regulatório (Mézard et al., 2007). Estes autores informam que, em todos os eucariotos estudados até aquele momento, incluindo plantas, a distribuição de *crossover* ao longo dos cromossomos

não é homogênea. Ou seja, a probabilidade local de ocorrência de um *crossover* varia com os diferentes intervalos em um cromossomo. Uma regra geral é o fato de que as regiões centroméricas são têm taxas de recombinação várias vezes menores que a taxa média num dado genoma.

A supressão da recombinação nas regiões centroméricas pode ser resultante do estado mais condensado da heterocromatina do centrômero na fase em que ocorrem os *crossover* durante a meiose, quando comparado com o estado de condensação da eucromatina. Porém, segundo Choo (1998), pesquisas em fungos usando centrômeros clonados indicaram que, mesmo centrômeros sem qualquer forma de heterocromatina provocaram diminuição significativa das taxas de recombinação nas adjacências das regiões onde os clones de centrômeros foram inseridos. Em *Drosophila*, a diminuição da recombinação nas regiões eucromáticas próximas a centrômeros é mantida mesmo após a retirada da heterocromatina centromérica. Parece que o efeito da supressão da recombinação é oriundo mais de uma atividade própria dos centrômeros do que da heterocromatina em si.

Williams et al. (1998) sugerem que a atividade dos centrômeros esteja relacionada com uma especificidade de organização mais elevada da cromatina dos centrômeros. Nesta linha, Warburton et al. (1997) encontraram uma proteína tipo histona, semelhante a H3, CENP-A (*conserved centromere associated protein – A*), que é específica aos centrômeros e que se associa somente a centrômeros ativos.

As plantas possuem grande variabilidade no tamanho, quantidade e estrutura dos cromossomos, de modo que para cada tipo de planta ocorre uma distribuição particular de *hotspots* e *coldspots* ao longo dos cromossomos. Em algumas espécies, incluindo trigo (*Triticum aestivum*), milho (*Zea mays*) e cevada (*Hordeum vulgare*), a taxa de ocorrência de *crossover* tende a aumentar com o aumento da distância física a partir do centrômero. Para o caso do milho, apresenta-se aqui uma ilustração (Figura 6), adapatada de Wang et al. (2006).

Com base nesse conjunto de observações, Mézard et al. (2007) sugeriram existirem vários níveis de controle da recombinação nas plantas, cada um operando numa escala diferente: genômica; cromossômica; regional (mega pares de bases) e local (milhares de pares de bases). Na escala genômica, a regulação pode se manifestar por meio do controle do número de quiasmas por cromossomo. Neste caso, são obedecidas duas regras principais: obrigatoriedade de ocorrência de pelo menos um quiasma por

cromossomo, para garantir sua segregação adequada; e, na maioria das espécies, o nível de intensidade da interferência. Ainda na escala genômica, a recombinação deve ser reprimida nas posições de longos fragmentos com sequências repetitivas, para proteger o genoma contra a possibilidade de ocorrência de grandes rearranjos.

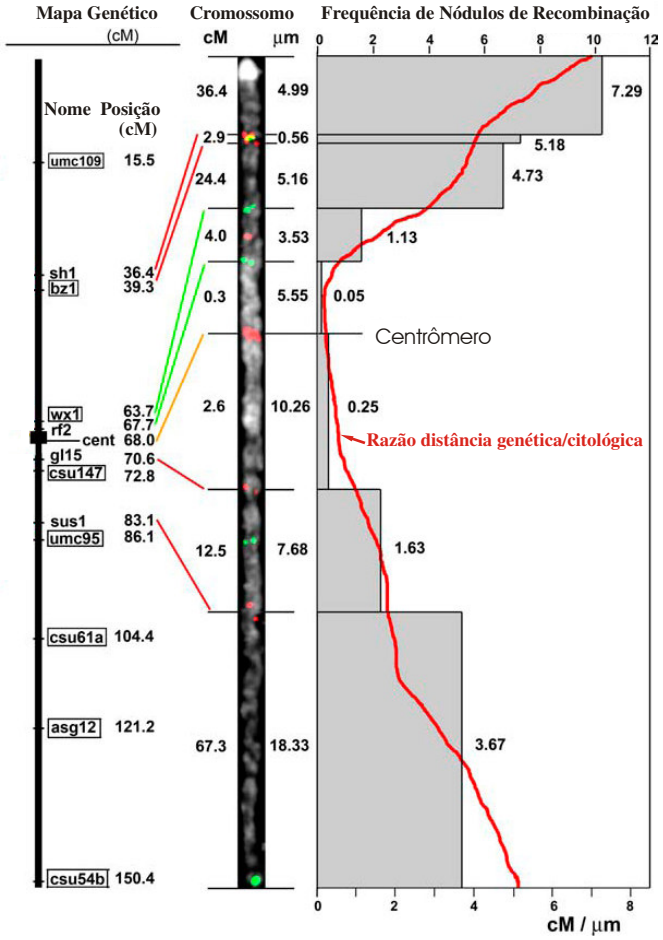


Figura 6. Mapa citogenético integrado do cromossomo 9 de milho. À esquerda: mapa genético redesenhado a partir do apresentado por Davis et al. (1999); as marcas estão em destaque. No centro: imagem do cromossomo 9; as diferenças nas cores foram convertidas em tons de cinza e sobrepostas com alguns sinais; à esquerda da imagem do cromossomo apresentam-se as distâncias genéticas (centimorgans) e, à direita, as distâncias citológicas (micrometros). À direita: a linha vermelha representa a razão entre distância genética por citológica (cM/µm); as barras em cinza representam o histograma da distribuição de freqüências de nódulos de recombinação estimada por Anderson et al. (2003). Fonte: Wang et al. (2006).

Drouaud et al. (2006) analisaram a variação da taxa de recombinação ao longo do cromossomo 4 de *A. thaliana*. Os autores verificaram que, mesmo dentro das regiões *hotspots* ou *coldspots*, a taxa de recombinação variava enormemente de um kilo par de bases para outro, conforme ilustrado na Figura 7.

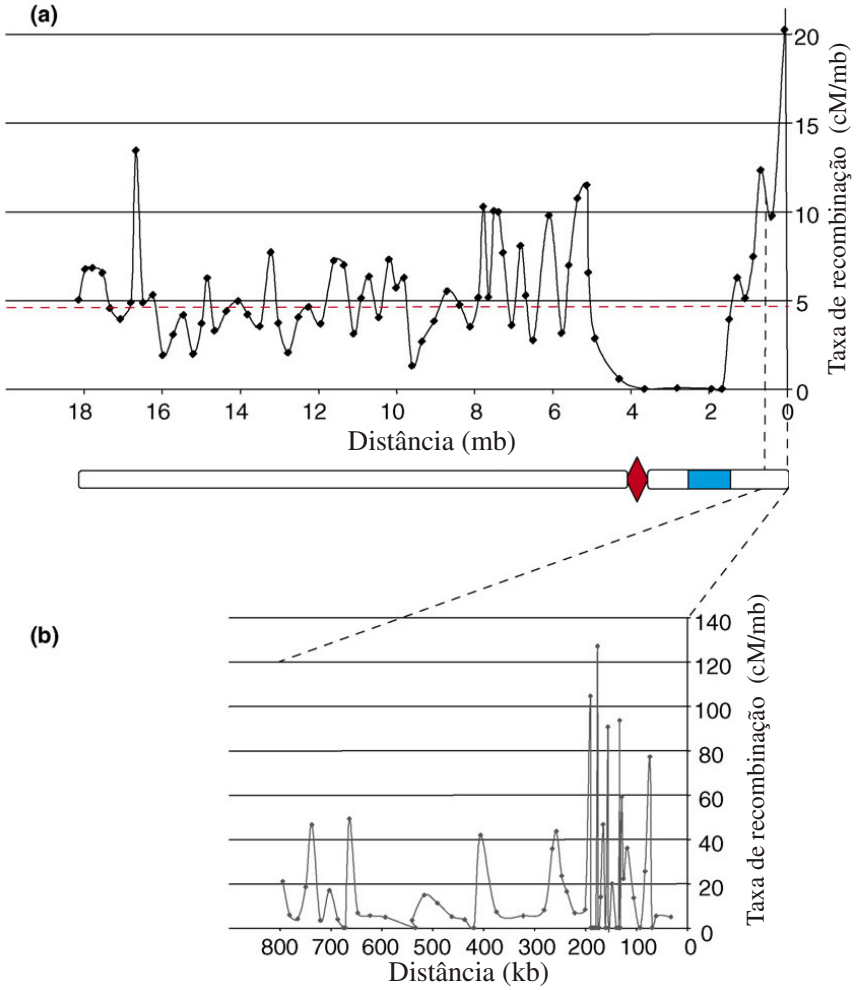


Figura 7. Variação da taxa de recombinação no cromossomo 4 de *A. thaliana*. a) A linha tracejada vermelha representa a média da taxa de recombinação no cromossomo 4 (4.6 cM/mb). Uma representação esquemática do cromossomo está alinhada com o gráfico. O retângulo ciano representa o *knob* heterocromático. O losango vermelho representa o centrômero. b) Variação na taxa de recombinação dentro dos primeiros 800 kb do braço menor do cromossomo, onde os pontos de ocorrência estão concentrados em dez regiões de tamanho menor que 5 kb, resultando em taxas de recombinação pelo menos cinco vezes maiores que a taxa média do cromossomo. Fonte: Drouaud et al. (2006).

A regulação em nível genômico, para garantir pelo menos um quiasma por cromossomo, pode ser ilustrada por meio da Figura 8. A presença de inúmeros nódulos de recombinação precoces ao longo de um cromossomo de tomateiro, na fase de zigóteno, evolui para a ocorrência de apenas um nódulo de recombinação tardio no final da fase paquíteno (Anderson & Stack, 2005).

Considerando a escala de cada cromossomo, em separado, o outro nível de regulação começa a agir: a presença ou não da interferência, e, se presente, a intensidade de sua ação ao longo do cromossomo.

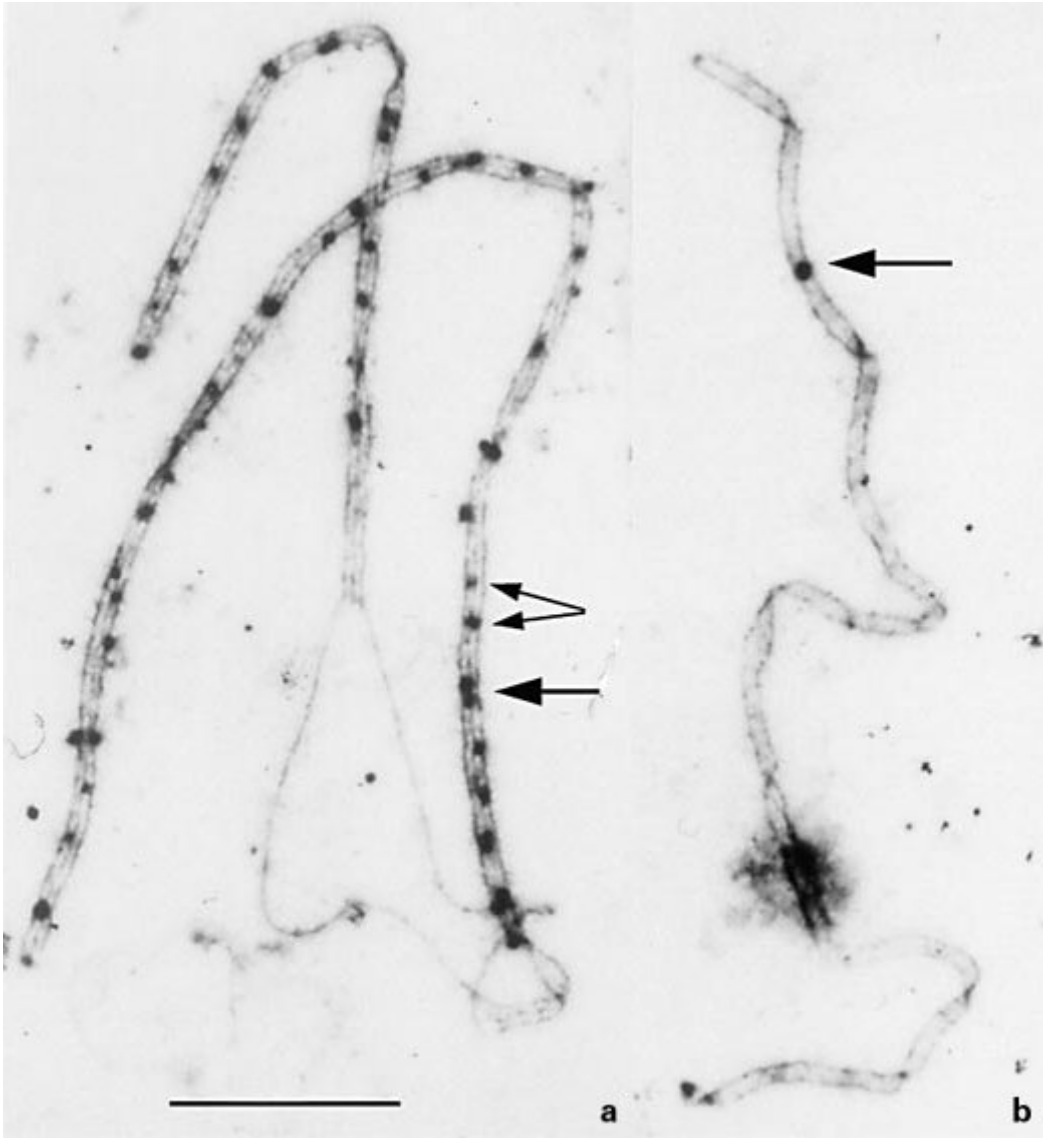


Figura 8. Imagens do complexo sinaptonêmico em célula de tomateiro. a) cromossomo na fase de zigóteno mostrando inúmeros nódulos de recombinação precoces ao longo das regiões em sinápse, e nenhum nódulo nas regiões sem sinápse mais ao centro. Os nódulos precoces variam em tamanho (dupla seta) e, às vezes, parecem se tocar (seta maior); b) cromossomo no final da fase paquíteno, com apenas um nódulo de recombinação tardio (seta). Fonte: Stack & Anderson (1986).

Mézard et al. (2007), com base na análise de estudos sobre mecanismos de regulação da distribuição de *crossover* e modelagem dos efeitos da interferência, sumarizaram uma opinião quanto à distribuição final dos pontos de *crossover* ao longo de um cromossomo. Segundo eles, o aspecto final da distribuição dos *crossover* ao longo de um cromossomo é resultado da integração de vários níveis de controle: *i*) a densidade de DSB varia ao longo dos cromossomos e a ocorrência de um DSB interfere na ocorrência de outros na vizinhança; *ii*) a probabilidade de um DSB ser resolvido como um *crossover*

provavelmente varia de região para região num cromossomo; *iii*) a interferência pode operar em vários níveis e sua intensidade varia ao longo do cromossomo e entre os cromossomos; *iv*) uma proporção do total de *crossover*, que varia de espécie para espécie, não sofre efeitos da interferência; e *v*) pode haver mais de duas ou três vias de resolução de DSB.

Esses diferentes níveis de controle de formação e resolução de *crossover* ainda são pouco entendidos. Não existe ainda uma teoria sistematizada sobre a diferenciação dos *crossover*, que tenha integrado os dados genéticos e citológicos, e que tenha proposto algum mecanismo molecular satisfatório para explicar a forma de distribuição dos *crossover* ao longo dos cromossomos. Cabe, porém, apresentar alguns exemplos que ilustram o padrão da variação da taxa de recombinação ao longo dos genomas em algumas espécies, e, em particular, apresentar a ocorrência de picos nas taxas de recombinação em fragmentos relativamente curtos de DNA, os denominados *hotspots* de recombinação.

Ao se observar o gráfico da Figura 7a, derivada dos estudos de Drouaud et al. (2006), observa-se uma região de aproximadamente 2 mb de extensão, entre o centrômero e a extremidade distal do *knob* heterocromático, com taxas de recombinação praticamente nulas. Essa quase total supressão da recombinação pode ser devido ao efeito conjunto da presença do centrômero e do *knob* heterocromático; mas, também há que se levar em consideração que esta é justamente a região que apresenta uma inversão, quando se compara a sequência de DNA do acesso Col com a do acesso Ler de *A. thaliana*. Estes mesmos autores mapearam 140 *crossover* em uma região de 800 kb localizada próxima à região organizadora de nucléolos, no braço menor do cromossomo 4 desta espécie. Eles observaram grandes variações na taxa de recombinação, com algumas sub-regiões apresentando total ausência de *crossover* (0 cM/Mb) e outras contendo número elevado de *crossover* (acima de 100 cM/Mb). Esse padrão de distribuição dos eventos, com grandes taxas de recombinação ocorrendo em pequenos fragmentos de DNA (Figura 7b), é bastante semelhante aos padrões de *hotspots* verificados em fungos, camundongos e seres humanos.

Paigen et al. (2008) relembram que, entre os mamíferos, a recombinação ocorre em regiões bem delimitadas denominadas *hotspots*, em geral com 1 kb a 2 kb de comprimento e com taxas de recombinação que podem variar até 1.000 vezes em sua intensidade. Os autores observaram também que as taxas de recombinação apresentam um bom grau de conservação em escala regional, mas não em escala local. Observaram, ainda, que em camundongos existia uma relação exponencial negativa entre o grau de atividade e

a abundância de *hotspots*, de modo que um pequeno número dos *hotspots* mais ativos é responsável pela maior proporção das recombinações. A Figura 9 apresenta o gráfico do padrão de variação da taxa de recombinação no cromossomo 1 de camundongo.

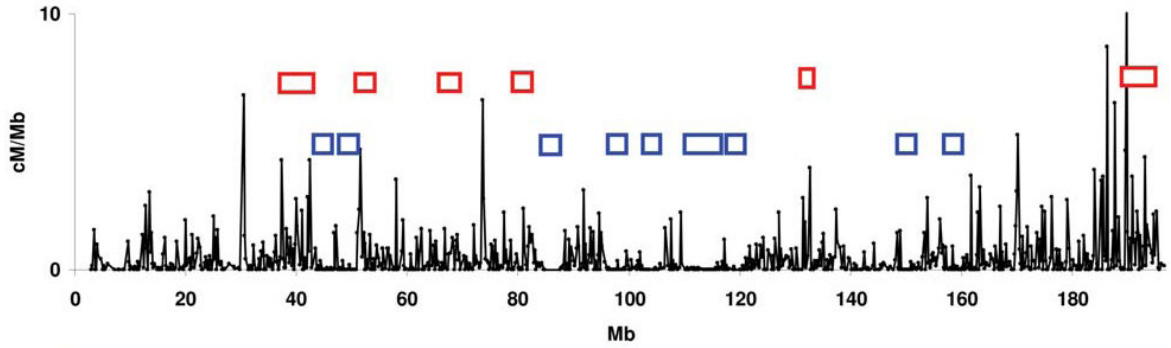


Figura 9. Taxa de recombinação ao longo do cromossomo 1 de camundongo. As pequenas caixas representam trechos com intervalos consecutivos de alta taxa de recombinação (vermelhos) ou com taxas nulas ou próximas de zero (azuis). Fonte: Paigen et al. (2008).

Yao et al. (2002) estudaram uma região de 140 kb no intervalo *a1-sh2* do genoma de milho, o qual contém pelo menos quatro gens (*a1*, *yz1*, *x1* e *sh2*). Eles mapearam a posição física de 101 pontos de recombinação e verificaram que estes não são uniformemente distribuídos no intervalo estudado, mas, contrariamente, são bastante concentrados dentro de três *hotspots* de recombinação. Dois desses *hotspots* estão dentro de áreas codificantes (*a1* e *yz1*) e o ouro não. A conclusão dos autores foi a de que nem todos os genes contêm *hotspots* e nem todos os *hotspots* estão dentro de genes.

Apesar disso, o que se prentende destacar aqui é o padrão da distribuição de *hotspots* de recombinação. Assim, numa escala mais local, o padrão consiste na ocorrência de grande quantidade de eventos de recombinação concentrada em regiões relativamente pequenas. Ou seja, grandes proporções dos eventos de recombinação ocorrem em pequenas proporções das sequências genômicas. A Figura 10 ilustra a ocorrência desse padrão no genoma humano.

Myers et al. (2006), usando análise *wavelet*, mediram o quanto da variação na taxa de recombinação é atribuída à variação em diferentes escalas físicas. Eles encontraram que a maior parte da variação da taxa de recombinação é dominada pela variação nas escalas entre 2 kb e 50 kb. Porém, a distribuição é bimodal, com significativa variação encontrada em escalas entre 2 mb a 30 mb, gerando um padrão de variação macro ao longo do cromossomo; ou seja, um padrão de variação com alcance espacial maior.

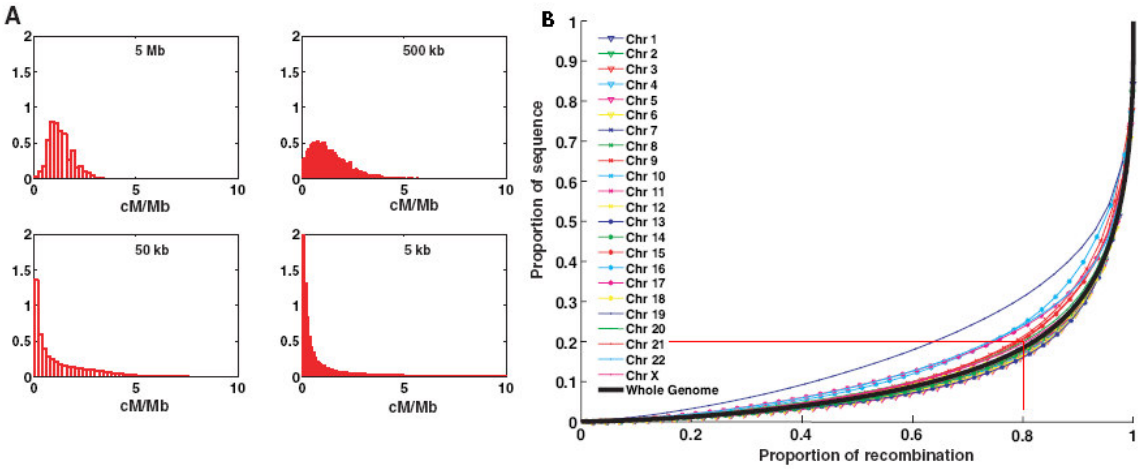


Figura 10. Variação da taxa de recombinação em escala local. A) Histogramas da taxa de recombinação para regiões de 5 mb, 500 kb, 50 kb e 5 kb ao longo do genoma, mostrando uma distribuição dominada por uma pequena proporção de grandes valores para as regiões menores; B) Proporção da recombinação total em várias porcentagens das sequências, plotada para cada cromossomo, mostrando que praticamente 80% de toda a recombinação ocorre em, aproximadamente, 20% das sequências. Fonte: Myers et al. (2005).

2.1.4 Elementos genômicos associados à variação da taxa de recombinação

Os estudos realizados por Jones et al. (2002) e Qi et al. (2002) mostraram que se a parte distal de um cromossomo de trigo, rica em genes, for extraída do cromossomo, a taxa de recombinação no segmento final dessa nova extremidade é bem maior do que a taxa que ocorria nesse mesmo segmento do cromossomo antes da excisão da parte distal. Mézard et al. (2007), comentando esses resultados, sugerem que a taxa de recombinação é determinada em maior parte pela localização de um fragmento de DNA do que pela sua sequência de nucleotídeos. Contudo, apesar deste comentário, muitos trabalhos têm sido realizados à procura de correlações entre a distribuição de elementos genômicos e a variação da taxa de recobinação, em vários organismos.

Myers et al. (2006) buscaram por essas correlações no genoma humano, usando uma série de elementos genômicos: localização de genes; composição de bases; densidade de elementos repetitivos; ilhas CpG; localização de duplicações; localização de segmentos conservados; localização de motivos associados a *hotspots* de recombinação e características epigenéticas como hipersensibilidade à DNase I. Apenas dois fatores tiveram boa correlação com a taxa de recombinação: o conteúdo GC (em escalas de 8 kb a 512 kb) e a localização do motivo associado a *hotspots* de recombinação (em escalas menores que 8 kb). Os autores ressaltam que, apesar de terem sido encontradas correlações

de alguns elementos genômicos com a taxa de recombinação, esses elementos têm baixo poder preditivo: todos os elementos com correlação significativa com a taxa de recombinação, juntos, explicam menos de 10% da variação total da taxa de recombinação.

Kendal & Suomela (2005) realizaram uma análise detalhada da distribuição de seis elementos genômicos nos cromossomos de *A. thaliana*: genes, indels, SNP, retrotransponsons, uma sequência repetitiva de 180 bp e uma classe de sequências centroméricas conservadas (CCS). Além da análise da distribuição intra-cromossomo, estes autores analisaram as correlações entre os seis elementos genômicos escohidos. Os resultados da análise espectral confirmaram a presença de um componente de variação de baixa frequência, de larga escala, nas flutuações de densidade dos seis elementos genômicos estudados. Essas flutuações tendem a se correlacionar umas com as outras em escalas genômicas mais maplas e nessas escalas as correlações tendem a ser mais fortes. Isso indica que as associações entre os diversos elementos são, provavelmente, de longo alcance espacial. Em função desses resultados, Kendal & Suomela (2005) propuseram que a estrutura cromossômica de maior escala tem uma influência dominante na distribuição dos elementos genômicos e proporciona um nível hierárquico na organização genômica em *A. thaliana*.

Marais et al. (2004) estudaram a associação entre a recombinação e o conteúdo GC e chegaram à conclusão de que esse elemento genômico não se correlaciona com as taxas locais de recombinação. Convém salientar que esses autores utilizaram, para estimar as taxas de recombinação ao longo dos cromossomos, os dados genéticos gerados por Wright et al. (2003), apenas 400 marcas em todo o genoma de *A. thaliana* e, talvez, essa escala de acesso às informações sobre a taxa de recombinação e ao conteúdo GC tenha impedido a verificação de qualquer provável correlação existente.

Já os autores Drouaud et al. (2006), usando uma maior densidade de marcas ao longo do cromossomo 4 de *A. thaliana*, encontraram correlação entre a variação da taxa de recombinação e elementos genômicos como o conteúdo GC e a densidade de alguns motivos de repetições de sequências simples (SSR).

A Figura 11 apresenta o padrão de distribuição dos elementos genômicos analisados por Kendal & Suomela (2005). A inspeção visual do padrão de distribuição dos elementos retrotransponsons, sequências repetitivas de 180 bp e as CCS já conduzem à intuição de que as correlações sejam significativas entre os elementos, pois estes obedecem a um mesmo padrão de distribuição, ou a padrões muito semelhantes entre si.

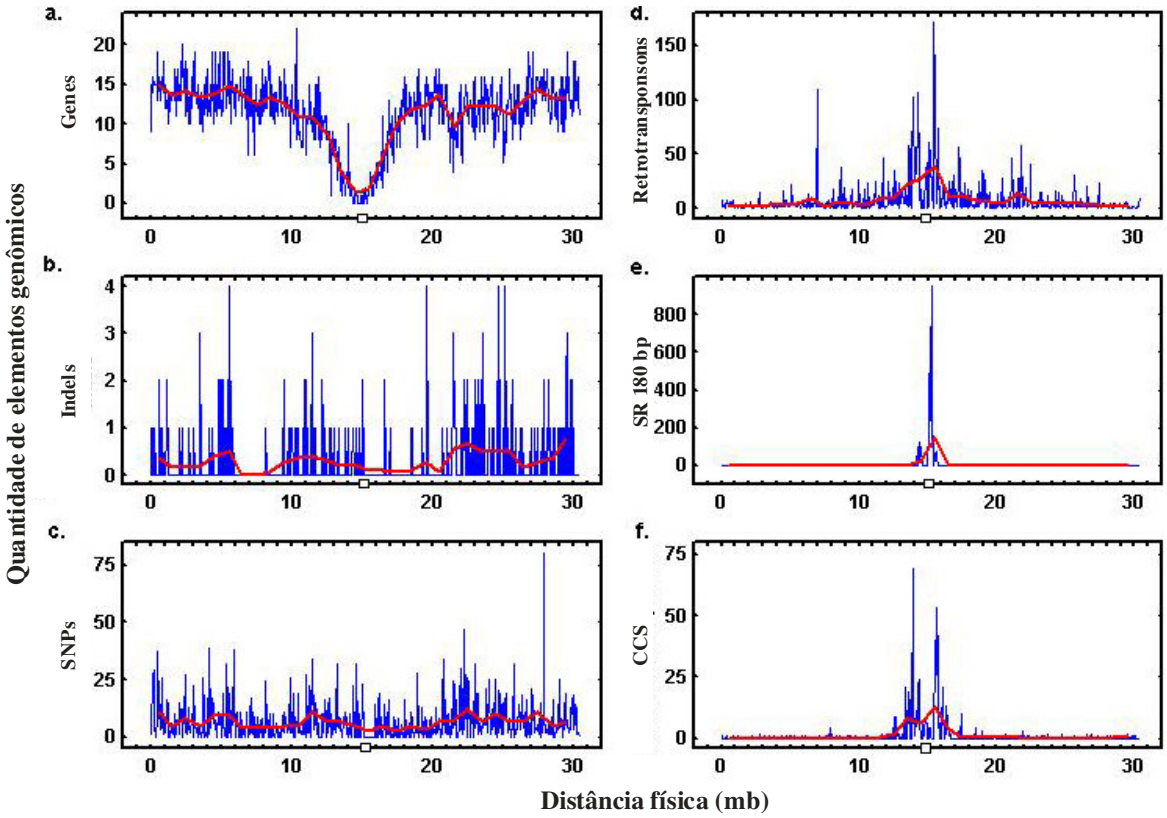


Figura 11. Padrões da distribuição de elementos genômicos ao longo do cromossomo 1 de *A. thaliana*. O número de elementos foi contado usando-se janelas de 50 kb (linhas azuis) e de 1 mb (linhas vermelhas). a) genes; b) indels: polimorfismos de inserções e deleções; c) SNPs: polimorfismos no nível de nucleotídeo; d) retrotransposons; e) SR 180 bp: sequências repetitivas de 180 bp; f) CCS: sequências centroméricas conservadas. A região centromérica está posicionada a aproximadamente 15 mb. Fonte: Kendal & Suomela (2005).

O gráfico da Figura 12 evidencia o fato de que as correlações entre os diversos elementos genômicos são dependentes da escala. Todas as correlações sofrem aumento do valor absoluto de seus coeficientes e de suas significâncias com o aumento da escala, como se observa a partir da escala de 200 kb. Exceto a correlação entre os elementos indel e SNP, todas as demais correlações sofrem uma queda acentuada a partir das escalas menores que 200 kb. Kendal & Suomela (2005) atribuem o rápido crescimento nos valores do coeficiente de correlação entre as escalas de 0 kb a 200 kb aos indícios de que um grande componente dessas correlações, em *A. thaliana*, pode ser atribuído a características cromossômicas de escalas maiores. Outros autores têm pesquisado à procura de correlações que se mantêm em escalas menores.

Chen & Zhao (2005) investigaram a hipótese de que a variação na taxa de recombinação está relacionada à organização estrutural de ordens mais elevadas do DNA. Para isso, estudaram a associação entre as variações locais na taxa de recombinação e os

níveis de simetria do DNA de camundongos, ratos e humanos. Para o cálculo das medidas de simetria eles se concentraram apenas nos oligonucleotídeos com 12 bases. Eles encontraram uma correlação negativa e significativa entre as taxas de recombinação e as simetrias complementares reversas nos genomas dos três organismos estudados. Essa correlação negativa se manteve quando os dados foram analisados para cada cromossomo em separado.

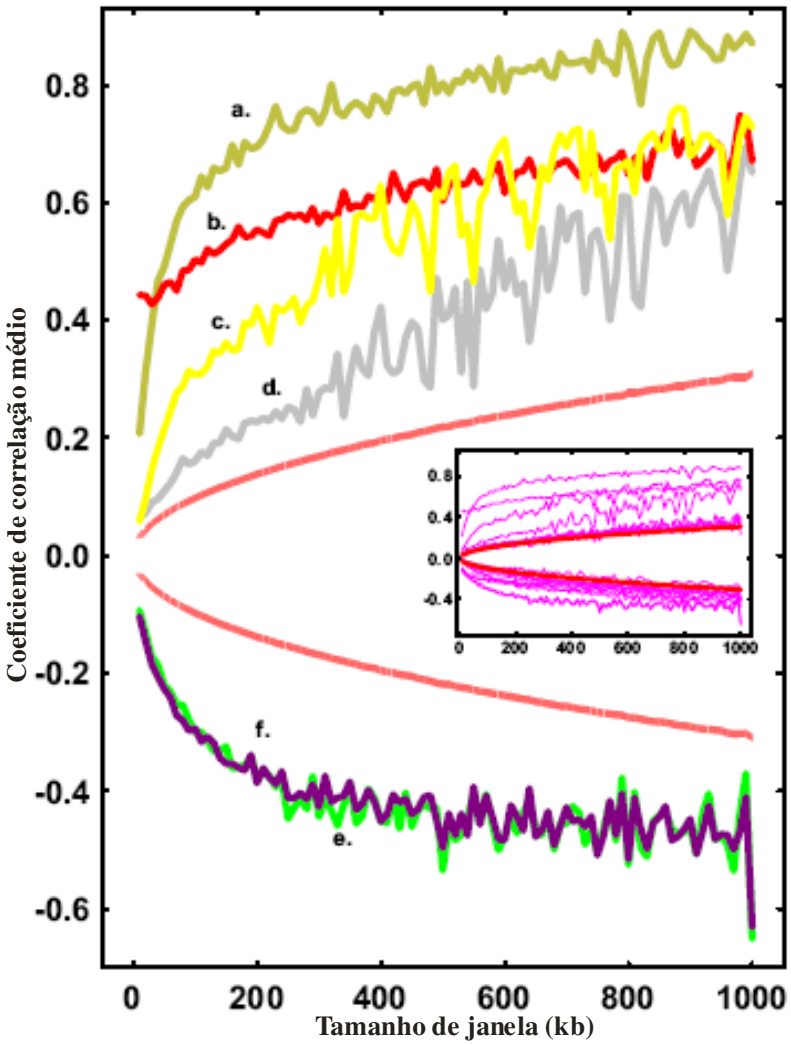


Figura 12. Coeficientes de correlação médios versus tamanho das janelas de contagem dos elementos genômicos. O coeficiente de correlação para os cinco cromossomos de *A. thaliana*, foi plotado contra os tamanhos das janelas usados para a contagem dos elementos genômicos. a) retrotransposons com sequências centroméricas conservadas; b) polimorfismos de inserções e deleções com polimorfismos no nível de nucleotídeo; c) retrotransposons com sequências repetitivas de 180 bp; d) sequências de 180 bp com sequências centroméricas conservadas; e) genes com retrotransposons; f) genes com sequências centroméricas conservadas. As linhas vermelhas centrais (obtidas por simulação) representam os valores críticos correspondentes a uma probabilidade de 5%. O quadro interno mostra as correlações para todas as 15 combinações estudadas. Fonte: Kendal & Suomela (2005).

Bagshaw et al. (2006) pesquisaram a associação entre trechos poli-purinas/poli-pirimidinas (poli-pu/poli-pi), uma classe de sequências com propriedades químicas distintas, e a recombinação em *hotspots*. A pesquisa foi realizada utilizando-se os dados de todo o genoma de *Saccharomyces cerevisiae* e dados de fragmentos de conhecidos *hotspots* de recombinação do genoma humano. A conclusão desses autores foi a de que uma possível relação funcional entre as sequências poli-pu/poli-pi e a localização e intensidade dos *hotsposts* de recombinação merece investigações futuras mais aprofundadas.

Myers et al. (2005) fizeram um amplo estudo da variação da taxa de recombinação e da localização de *hotspots* de recombinação ao longo do genoma humano. Foram identificados mais de 25 mil desses *hotsposts*, alguns motivos de sequências e contextos que desempenham certo papel na atividade dos *hotsposts* de recombinação. Os autores investigaram a freqüência de todas as combinações possíveis de sequências com cinco a nove nucleotídeos em todo o genoma. Dentre as 8.192 sequências com sete bases, o motivo CCTCCCT apresentou uma freqüência de 5,1 a 5,7 vezes maior dentro das regiões de *hotspots* do que no resto do genoma ($p<10^{-33}$ a $p<10^{-5}$).

Nessa tentativa de encontrar uma sequência que possa estar associada à localização ou ao grau de atividade dos *hotsposts* de recombinação existem vários outros trabalhos. Blumental-Perry et al. (2000) identificaram o motivo que eles denominaram de CoHR (*Common Homology Region*) e o consideraram como bom preditor das localizações das lesões meióticas DSB. Kirkpatrick et al. (1999) encontraram evidências de que a sequência $(CCGNN)_{12}$ elevava a atividade de recombinação nos *hotspots*, enquanto a sequência $(CCGNN)_{48}$ diminuia essa atividade, em fungos. Steiner & Smith (2005a) identificaram uma sequência de sete bases, localizada no *hotspot* de recombinação M26, em *S. pombe*, que determina a distribuição e, portanto, tem boa predição para a localização de *hotspots* de recombinação meiótica. Esse heptâmero, 5'-ATGACGT-3', consiste num ponto de ligação do fator de transcrição Atf1-Pcr1. Em um trabalho subseqüente, Steiner & Smith (2005b) verificaram que a sequência consenso que melhor descreveria a posição de ligação do fator de transcrição Atf1-Pcr1 consistia numa ampliação da sequência M26, passando de sete para dezoito pares de bases: 5'-GNVTATGACGTCATNBNC-3'.

Em vários estudos sobre possíveis associações de elementos genômicos ou de motivos de sequências com a recombinação, os autores hipotetizam sobre a influência de fatores estruturais do DNA sobre a variação da taxa de recombinação ao longo dos

cromossomos. Porém, quantificar ou conseguir estimativas numéricas sobre a variação de uma propriedade intrinsecamente estrutural não é tarefa fácil. Para contribuir com o preenchimento desta lacuna, Greenbaum et al. (2007) idealizaram e estabelecram um método de obter estimativas de uma informação de natureza estrutural em cada nucleotídeo de uma sequência. O padrão de clivagem do DNA por radicais hidroxila é uma medida da variação local da acessibilidade do solvente à superfície da dupla fita de DNA e, portanto, oferece uma informação local sobre a forma e estrutura do DNA. Então, esses autores produziram uma quantidade extensa de dados de clivagem por radical OH, usando duas bibliotecas de DNA. Esses dados foram armazenados em um banco de dados relacional. Também foram desenvolvidos algoritmos capazes de predizer o padrão da intensidade de clivagem por radicais OH, numa sequência de DNA, com grande precisão. Uma descoberta interessante dos autores foi a constatação de que fragmentos de DNA bastante distintos, quanto às suas sequências, comungavam um mesmo padrão estrutural.

Alguns autores, Breyne et al. (1994) e Nabirochkin et al. (1998), postulam a existência de algumas sequências específicas de DNA que mediam a ligação dos cromossomos à matriz nuclear e/ou ao citoesqueleto nuclear. As primeiras são, geralmente, denominadas de regiões de ligação à matriz nulcear, ou *Matrix Attachment Regions* – MAR. As últimas denominadas de regiões de ligação ao citoesqueleto nuclear, ou *Scaffold Attachment Regions* – SAR. Em geral, quando os autores realizam buscas genômicas tentando identificar elementos destes dois tipos de sequências, ou regiões de sequências, os resultados são apresentados com a denominação comum de elementos MAR/SAR, indicando a provável ocorrência de sequências de ambas as classes. Breyne et al. (1994) informam haver indícios de que sequências do tipo MAR/SAR sejam responsáveis pela organização estrutural e funcional dos genomas das plantas. Enquanto alguns tipos de elementos MAR/SAR delimitam loops estruturais de cromatina, flanqueando unidades funcionais de expressão gênica e replicação do DNA, outros tipos se localizam próximos a elementos regulatórios, podendo estarem diretamente envolvidos na regulação da expressão gênica.

Para verificar se há alguma associação entre a recombinação e alguma informação de natureza estrutural, nas análises do presente estudo, será verificado se as variações na taxa de recombinação têm alguma associação com a variação da intensidade de clivagem por radicais OH^- ou com a distribuição de elementos MAR/SAR.

2.2 MÉTODOS PARA ESTIMAR A TAXA DE RECOMBINAÇÃO USANDO DADOS GENÉTICOS DE POPULAÇÕES

2.2.1 Fatores que afetam o processo de recombinação

Antes de adentrar nos detalhes de como obter estimativas sobre taxas de recombinação, usando sequências de DNA amostradas de populações, e qual modelo estatístico usar para obter estas estimativas, faz-se necessária a contextualização sobre os vários fatores que afetam o processo de recombinação. Para isso, foi escolhida parte da revisão elaborada por Stumpf & McVean (2003) e a descrição apresentada por Allard (1963).

Segundo Stumpf & McVean (2003), quando uma nova mutação ocorre em determinado ponto de um cromossomo, o novo alelo que surge estará em completa associação com todos os outros pontos polimórficos presentes naquele cromossomo. Ao longo do tempo, essas associações são desfeitas pela ação do processo de recombinação, de modo que, teoricamente, o grau do desequilíbrio de ligação entre os alelos é, simplesmente, uma função da taxa de recombinação e da idade de ocorrência do evento de mutação. Porém, outros fatores evolutivos também afetam o padrão do desequilíbrio de ligação: a história da população (mudanças no tamanho da população, estruturação geográfica), mutação, seleção natural, deriva genética.

Por outro lado, as estimativas da taxa de recombinação entre dois locos quaisquer podem variar em função de fatores ambientais e genéticos. Allard (1963) investigou a variação das estimativas da taxa de recombinação em função do material genético e de fatores ambientais, em linhagens puras de *Phaseolus lunatus* e híbridos F1 de cruzamentos entre as linhagens. Numa parte da experiência, visando avaliar os efeitos de ambientes nos valores das taxas de recombinação, plantas F1 de quatro cruzamentos foram plantadas em dois anos consecutivos, em três épocas distintas em cada ano. Em outra parte da pesquisa foi praticada a seleção para altas e baixas taxas de recombinação entre progênies autofecundadas de alguns híbridos, para avaliar o grau de controle genético da recombinação. Os locos utilizados para estimar a taxa de recombinação eram sabidamente ligados e de fácil avaliação. A magnitude da variação na porcentagem de recombinação entre as famílias F1 dentro das épocas de plantio foi de quase seis pontos percentuais, enquanto essa magnitude chegou a 17 (de 27% para 44%) pontos percentuais entre as diferentes épocas. O *background* genético também teve efeito significativo na variação da

percentagem de recombinação. Em cada planta F1 as sementes eram colhidas à medida que chegavam à maturidade e, em seguida, a porcentagem de recombinação era avaliada para cada lote separadamente. Verificou-se que a porcentagem de recombinação era relativamente homogênea dentro de cada época de colheita, mas, bastante variável entre as diferentes épocas. No experimento para baixas e altas taxas de recombinação, cinco plantas F2 de cada híbrido foram escolhidas ao acaso e, em seguida, a taxa de recombinação foi determinada para cada planta, a partir dos dados de segregação de suas respectivas progênies F3. De cada progênie F3, cujos genitores F2 apresentaram as maiores taxas de recombinação, eram escolhidas três plantas ao acaso. Da mesma forma foi feito para as plantas F3 cujos genitores F2 apresentaram as menores taxas de recombinação. Esse processo foi repetido até a geração F6. As linhagens selecionadas mostraram uma flutuação ano a ano na taxa de recombinação. Essa flutuação era semelhante à observada entre épocas de plantio para os híbridos F1, e não foi suficiente para camuflar a forte tendência apresentada na resposta à seleção no sentido das taxas de recombinação mais altas. As médias da taxa de recombinação entre linhagens F5-alta e F5-baixa, medidas pelas proporções de segregação na geração F6, mostraram que as progênies selecionadas para altas taxas de recombinação eram de 7,5 a 20 vezes mais recombinogênicas do que as selecionadas para baixas taxas de recombinação. O ganho de seleção foi quase que completamente na direção do aumento de recombinação; o ganho no sentido de baixas taxas de recombinação ocorreu, mas foi pequeno. Supondo-se que as mudanças foram decorrentes de genes que afetam a taxa de *crossover*, a estabilização do valor da taxa de recombinação no sentido de sua redução sugere que mais de um (ou vários) gene está envolvido nesse controle. O rápido ganho no sentido de aumentar a taxa de recombinação sugere que alguns genes podem ter tido grandes efeitos. Isso mostra que o processo de recombinação está sob controle genético e que as taxas de recombinação sofrem influências significativas de fatores ambientais e relacionados ao *background* genético.

Diante do exposto, percebe-se a complexidade presente nas tentativas de estimar taxas de recombinação e caracterizar sua variação ao longo de um genoma. A magnitude das variações e da aleatoriedade que caracterizam os fatores evolutivos torna proibitivo o uso de modelos simples e determinísticos. Portanto, extrair o sinal da variação da taxa de recombinação ao longo de um genoma, a partir de dados genéticos de populações, representa um grande desafio, tanto do ponto de vista estatístico quanto computacional (Stumpf & McVean, 2003).

2.2.2 Taxa de recombinação populacional e taxa de recombinação por geração

Dentro do contexto dos métodos apresentados nas partes subsequentes da presente revisão, o parâmetro-chave considerado como modelador do padrão de variação de desequilíbrio de ligação ao longo de um genoma é a taxa de recombinação populacional ρ, estimado por $\rho = 4N_e r$, em que N_e é o tamanho efetivo populacional e r é a taxa de recombinação por geração. A taxa de recombinação por geração pode depender de fatores genômicos como motivos de sequências de bases, propriedades estruturais do DNA, conteúdo GC, entre outros. Já a taxa de recombinação populacional pode depender também de fatores demográficos, relacionados ao tamanho efetivo populacional, e, portanto, pode variar bastante entre populações.

2.2.3 Métodos para estimar a taxa de recombinação populacional

Há inúmeros métodos que têm sido utilizados para acessar informação sobre a taxa de recombinação populacional. Para Stumpf & McVean (2003), estes podem ser enquadrados em, pelo menos, cinco grandes categorias: métodos baseados na contagem de eventos de recombinação; métodos baseados no modelo de coalescência; métodos que usam estatísticas descritivas; métodos que usam funções de verossimilhança; e métodos que usam aproximações às funções de verossimilhança.

2.2.3.1 Métodos baseados na contagem de eventos de recombinação

Os métodos não baseados em modelagem de processos se concentram em contabilizar o número de eventos de recombinação que ocorreram na história de uma amostra. Esses métodos supõem que os sucessivos eventos de recombinação, que ocorreram ao longo da história de uma amostra, deixam uma assinatura ou um padrão identificável por meio do uso de dados genéticos de populações. Dois deles foram propostos por Hudson & Kaplan (1985) e Myers & Griffiths (2003), sendo ambos fundamentados num método apresentado por Weir (1979), denominado teste dos quatro gametas (ou *Four-Gamete Test - FGT*). A estatística R_m de Hudson & Kaplan (1985) fornece estimativas conservadoras do número de eventos de recombinação e, em geral, subestima em muito o número mais provável de eventos ocorridos em um intervalo do

genoma. Já a estatística R_h, do método de Myers & Griffiths (2003), faz uso da comparação entre o número de haplótipos (*M*) e o número de pontos polimórficos (*N*). Se *M* haplótipos são observados em uma região com *N* marcas polimórficas, então *M-N* eventos de recombinação devem ter ocorrido naquela região (se *M<N*, considera-se que o número de eventos de recombinação é zero). Um terceiro método não-paramétrico para contabilizar o número mínimo de eventos ocorridos numa região genômica foi elaborado por Wiuf (2002). Trata-se de um método sofisticado, mas tecnicamente complexo. Stumpf & McVean (2003) ressaltam que esses três métodos propiciam estimativas subestimadas do verdadeiro número de eventos de recombinação ocorridos numa região genômica, e atribuem isto ao fato de que são necessárias condições específicas para que seja possível detectar eventos de recombinação. Essas condições estão relacionadas com a diversidade genética na região em estudo, a idade do evento de recombinação, o tamanho da amostra e a história demográfica da população.

Além de ser quase impossível contar todos os eventos de recombinação, pelo fato de que nem todos deixam suas marcas no genoma, outro problema é a determinação do número de gerações em que ocorreram esses eventos. Só assim seria possível estimar a taxa de recombinação por geração. Assim, Stumpf & McVean (2003) sugerem que os métodos que modelam explicitamente os processos em estudo podem superar estas dificuldades. Sugerem também que os métodos baseados no modelo de coalescência podem ser considerados mais adequados, por tratarem-se de modelos probabilísticos, que descrevem as distribuições da genealogia subjacente a uma amostra de cromossomos em uma população idealizada.

2.2.3.2 Métodos baseados no modelo de coalescência

Wakeley (2006) esclarece que coalescer significa crescer junto, juntar ou fundir. Quando duas cópias de um gene descendem de um mesmo ancestral comum em alguma geração do passado, diz-se que estas cópias se coalescem naquela geração. Kingman (1982) informa que a coalescência de duas cópias de um gene em um ancestral comum pode ser descrita por um processo modelado matematicamente, ao qual ele denominou de processo *n*-coalescente. O processo de coalescência abordado no contexto de modelos de população, como o modelo de Wrigt-Fisher e o de Moran, consiste na base dos métodos mais recentes de estimação da taxa de recombinação a partir de dados

genéticos populacionais.

Li & Stephens (2003) ressaltaram que algumas abordagens têm obtido maior êxito na construção de modelos estatísticos que relacionam a variação presente em dados genéticos de amostras populacionais com a taxa de recombinação. Estas abordagens são baseadas no modelo de coalescência, proposto por Kingman (1982), e em suas generalizações para incluir a modelagem do processo de recombinação, como a proposta apresentada por Hudson (1983).

As versões ou extensões do modelo de coalescência para incluir a modelagem da recombinação são coletivamente denominadas *ARG*, de *Ancestral Recombination Graph*. Estes modelos compartilham as seguintes suposições: tamanho populacional constante, cruzamentos ao acaso, evolução com base na teoria da neutralidade, taxa de recombinação uniforme ao longo do genoma.

2.2.3.3 Métodos que usam estatísticas descritivas

Alguns autores, visando fazer inferências sobre a taxa de recombinação populacional (ρ), construíram estimadores que usam a informação contida em algumas estatísticas descritivas dos dados. Por exemplo, a variância das diferenças entre pares de sequências de nucleotídeos pode ser interpretada como uma medida de desequilíbrio de ligação. Então, essa variância ou outra medida descritiva pode ser usada para inferir sobre ρ. Como esclarecem Stumpf & McVean (2003), esses estimadores são fáceis de calcular e exigem pouco esforço computacional, mas eles não incluem toda a informação que está contida nos dados genéticos de populações. Nesta categoria incluem-se o método de Hudson (1985) e o método de Wall (2000). Uma comparação entre vários métodos que fazem uso de estimadores baseados em estatísticas descritivas pode ser encontrada no trabalho de Wall (2000). Este autor reconhecia que os métodos baseados em verossimilhança tinham a vantagem de usar toda a informação contida nos dados, porém, a desvantagem de serem computacionalmente intensivos.

2.2.3.4 Métodos que usam funções de verossimilhança

Stumpf & McVean (2003) descrevem que os métodos baseados em funções de verossimilhança estimam a probabilidade de se observar um determinado conjunto de

dados sob as suposições de um determinado modelo genético de população. Esses métodos incorporam a estrutura genealógica subjacente aos dados de uma amostra e tentam usar a totalidade das informações contidas nos dados. São métodos muito exigentes em recursos computacionais. Em geral, o parâmetro estimado é a taxa de recombinação populacional, mas os modelos podem incorporar taxa de mutação e parâmetros demográficos. Por meio desses métodos são realizadas intensas e extensas simulações para estimar a superfície de verossimilhança para os parâmetros do modelo escolhido. A verossimilhança de um parâmetro, dado um conjunto de observações, é proporcional à probabilidade deste conjunto dado os parâmetros do modelo. Por exemplo, o valor de ρ para o qual é máxima a probabilidade de se observar o conjunto de dados é considerada a estimativa de máxima verossimilhança de ρ.

A complexidade do cálculo de verossimilhanças em genética de populações cresce rapidamente com o tamanho do conjunto de dados. Como as expressões analíticas são impossíveis de se derivar, a não ser para pequenos conjuntos de dados, então, o cerne dos métodos baseados em funções de verossimilhança está nas técnicas utilizadas para realizar as simulações. A idéia central das técnicas de simulação (ou Monte Carlo – MC), de acordo com Stumpf & McVean (2003), "é ampliar ou expandir o conjunto de dados original incluindo todas as genealogias possíveis, considerando os processos de coalescência, de recombinação e, em geral, de mutação. Com o conjunto de todas as genealogias possíveis, pode-se calcular a probabilidade de se observar o conjunto de dados originais dado o conjunto de genealogias, ou calcular a probabilidade de determinada genealogia, dado o modelo coalescente e seus parâmetros. Como podem existir, virtualmente, infinitos conjuntos de genealogias que podem ter originado os dados, encontrar aquelas que são mais prováveis, considerando o modelo adotado, é comparável ao provérbio da "agulha no paiol".

Para contornar o problema mencionado acima, duas estratégias de simulação têm sido mais amplamente utilizadas para realizar inferências em genética de populações: a denominada *Importance Sampling* (*IS*) e outra baseada em cadeias de Markov, ou *Markov Chain Monte Carlo* (*MCMC*). Dentro de cada uma dessas estratégias existem várias alternativas de algoritmos.

Na estratégia *MCMC* arbitra-se um ponto inicial e, em seguida, fazem-se transições ou mudanças para outros pontos da genealogia. As mudanças que são mais prováveis são aceitas e as menos prováveis, rejeitadas, com uma probabilidade que é

proporcional à possibilidade de se observar esse novo ponto, dado o modelo em uso. De acordo com Stephens (2007), os métodos baseados na estratégia *IS* vieram para aumentar a eficiência das integrações presentes na estratégia *MCMC*, pois, como a designação sugere, os esforços de computação são concentrados apenas no subconjunto das genealogias simuladas que é mais consistente com os dados observados.

Os pioneiros no uso dos algoritmos *IS* no contexto de genética de populações foram Griffiths & Tavaré (1994a, 1994b, 1994c), que derivaram um método para solucionar sistemas de equações recursivas. Ainda segundo Stephens (2007), foram Felsenstein et al. (1999) que observaram a conexão existente entre o método de Griffiths & Tavaré e os algoritmos *IS*, e, em seguida, Stephens & Donnelly (2000) demonstraram como essa observação pode ser explorada para desenvolver algoritmos ainda mais eficientes.

Na categoria dos métodos baseados em funções de verossimilhança citam-se: Griffiths & Marjoran (1996), Kuhner et al. (2000), Nielsen (2000), Fearnhead & Donnelly (2001) e Wang & Rannala (2008).

2.2.3.5 Métodos que usam aproximações às funções de verossimilhança

Como os métodos baseados em funções de verossimilhança são computacionalmente bastante intensivos, mesmo para conjuntos de dados relativamente pequenos, foram desenvolvidos vários métodos que se baseiam em aproximações às funções de verossimilhança, visando aplica-los em conjuntos de dados mais amplos, incluindo cromossomos ou genomas completos. Dentre estes, destacam-se os métodos de Hudson (2001), Fearnhead & Donnelly (2002), McVean et al. (2002) e Li & Stephens (2003).

Stumpf & McVean (2003) informam que esses métodos usam duas táticas principais: ou ignoram as marcas com alelos de baixa frequência na população, as quais carregam pouca informação sobre os eventos de recombinação; ou dividem o total de marcas em subconjuntos, utilizando um subconjunto de cada vez. Numa tática ou na outra calculam-se verossimilhanças parciais para cada subconjunto de dados e, em seguida, as verossimilhanças parciais são combinadas para obter um estimador da verossimilhança composta. A estimação por verossimilhança composta consiste numa razoável aproximação à estimação por máxima verossimilhança para o conjunto completo de dados.

A subdivisão pode chegar em nível de um par de locos, para o qual é considerada a distribuição alélica de um sistema de dois locos. Então, para cada par de locos é calculada, de forma independente, uma superfície de verossimilhança para os parâmetros de interesse (taxa de recombinação, mutação, etc.). Uma verossimilhança composta é calculada por meio do produtório de todas as verossimilhanças parciais para pares de locos. Segundo Stumpf & McVean (2003), apesar de que esses métodos ignoram muito da informação contida nos dados, eles são rápidos e eficientes do ponto de vista computacional. Em termos de viéses e variâncias, estudos com simulações mostraram que as estimativas obtidas com esses métodos têm acurácia semelhante à de outros métodos desenvolvidos para a mesma finalidade.

Outras vantagens dos métodos de verossimilhança composta são: *i*) o número das possíveis combinações de alelos num sistema de dois ou poucos locos pode ser facilmente calculado, mesmo sem referência direta aos dados e, depois, as respectivas tabelas podem ser utilizadas para aumentar a eficiência computacional do processo de cálculo das superfícies de verossimilhança; *ii*) dados genotípicos podem ser usados diretamente, sem a necessidade de inferir os haplótipos; *iii*) é possível modelar situações mais complexas de mutação e de modelos populacionais.

Vários outros métodos baseados em aproximações às verossimilhanças foram desenvolvidos. Cada um tenta captar os aspectos mais informativos da história genealógica de uma amostra (*ARG*), para estimar a taxa de recombinação. Nota-se uma tendência de os novos métodos incorporarem modelos demográficos, novos modelos do processo de recombinação, bem como modelar as influências dos delineamentos experimentais.

2.2.4 Comparações entre métodos de estimação da taxa de recombinação populacional

Uma comparação entre os métodos de estimação da taxa de recombinação populacional deve levar em consideração aspectos como: viés, variância, consistência, robustez diante de desvios em relação às suposições do modelo assumido, eficiência computacional. A literatura científica está repleta de trabalhos que comparam os diversos métodos disponíveis para tal. Em geral, ao apresentar um novo método, os autores já fazem uma comparação com outros métodos semelhantes.

Stumpf & McVean (2003) resumem, de forma simples, a visão geral que se

deve ter com relação às principais diferenças entre os métodos: “Os métodos baseados em contagens de eventos de recombinação são mais rápidos do ponto de vista computacional, mas não usam a totalidade das informações contidas nos dados; têm considerável variância e podem ser bastante afetados por pequenos desvios em relação às suposições do modelo adotado. Os métodos que usam funções de verossimilhança, por outro lado, fazem uso mais pleno das informações contidas nos dados, mas são computacionalmente muito intensivos e, às vezes, até proibitivos. Os métodos baseados em aproximações às funções de verossimilhança estão entre os dois extremos, ressaltando que alguns ainda carecem de consistência. Aqueles que são baseados em distribuições de sistema de pares de locos parecem ser menos influenciados por efeitos demográficos que os baseados em múltiplos locos, e também por problemas relacionados à genotipagem de SNP”.

2.2.4.1 O método de Fearnhead & Donnelly (2001)

Fearnhead & Donnelly (2001) introduziram um método para estimar a taxa de recombinação usando dados genéticos de populações, tendo como base o modelo de coalescência. O método inclui a modelagem da recombinação e trata-se de um procedimento estatístico computacionalmente intensivo, usando a abordagem baseada em máxima verossimilhança completa. A extensão do modelo coalescente para incluir a recombinação é denominada de gráfico das recombinações ancestrais (*Ancestral Recombination Graph - ARG*). Os autores propõem obter uma aproximação para a superfície de verossimilhança conjunta das taxas de recombinação e de mutação. Essa superfície envolve a soma sobre todas as possíveis genealogias consistentes com os dados. Como a avaliação exata dessa soma é praticamente impossível, eles desenvolveram um procedimento baseado no algoritmo *IS* para aproximar essa superfície. Os autores compararam o novo método com aquele anteriormente apresentado por Griffiths & Marjoram (1996) e com de *Markov Chain Monte Carlo* (MCMC), proposto por Kuhner et al. (2000). A comparação se deu no âmbito da acurácia da aproximação da superfície de verosimilhança e das propriedades dos estimadores das taxas de recombinação e de mutação. Fearnhead & Donnelly (2001) concluíram que o método por eles desenvolvido mostrou-se substancialmente mais eficiente que os outros, em ambos os aspectos da comparação.

Houve desenvolvimento significativo do método de Fearnhead & Donnelly

(2001), em relação aos de Wall (2000), Griffiths & Marjoram (1996) e Kuhner et al. (2000). Porém esses métodos tiveram desempenho satisfatório apenas quando aplicados a conjunto de dados de regiões relativamente pequenas do genoma humano. Para regiões do genoma com tamanhos um pouco mais elevados, esses métodos já se tornam computacionalmente impraticáveis. Visando reduzir a alta demanda por recursos computacionais, outros autores propuseram o uso de métodos baseados em aproximações às funções de verossimilhança, seguidas do cálculo de uma verossimilhança composta (Hudson, 2001; McVean et al., 2002; Fearnhead & Donnelly, 2002).

2.2.4.2 O método de Hudson (2001)

O método de Hudson (2001) é baseado em aproximações às verossimilhanças. Verossimilhanças para pares de locos bi-alélicos, são obtidas via simulação, assumindo o modelo mutacional de alelos infinitos, ou seja, sem mutações recorrentes. O autor concentra sua atenção nos casos em que a taxa de mutação é muito pequena, caso particular em que o modelo de alelos infinitos se assemelha bastante ao modelo de pontos infinitos. Porém, é sabido que o processo de mutação recorrente pode produzir um padrão de diversidade genética semelhante ao produzido pelo processo de recombinação, causando viéses nas estimativas da taxa de recombinação populacional.

2.2.4.3 O método de McVean et al. (2002)

O método de McVean et al. (2002) é uma extensão ao método de Hudson (2001) para construir um modelo que permite altas taxas de mutações recorrentes.

2.2.4.4 O método de Fearnhead & Donnelly (2002)

Fearnhead & Donnelly (2002), ao invés de desenvolverem um método para obter uma aproximação da verossimilhança completa, optaram por considerar duas verossimilhanças distintas. A primeira, uma verossimilhança marginal pela qual alguma informação contida nos dados é desprezada, obtendo-se, então, uma verossimilhança para um subconjunto dos dados. Esta aproximação é semelhante ao método proposto por Wall (2000). Adotando-se uma cuidadosa escolha sobre quais aspectos dos dados poderiam ser

ignorados, eles obtiveram considerável redução do serviço computacional às custas de pequena perda de informação para estimar os parâmetros de interesse. A segunda aproximação usa uma modelagem mais simples para os dados, ignorando as correlações de longo alcance. Essa idéia tem sua fundamentação em alguns aspectos da estatística espacial. Ao invés de compor uma log-verossimilhança composta pela soma das log-verossimilhanças de todos os pares de locos, Fearnhead & Donnelly (2002) dividem a região de interesse em sub-regiões e compõem a log-verossimilhança composta, somando-se as log-verossimilhanças de todas as sub-regiões. A acurácia das inferências baseadas neste método depende do tamanho adotado para as sub-regiões. Quanto maior a sub-região mais próxima a log-verossimilhança composta estará da log-verossimilhança completa (ótima).

Nenhum dos métodos até aqui apresentados se preocupam em modelar a possibilidade de ocorrer variação na taxa de recombinação ao longo da região de interesse. Todos trabalharam com a obtenção de estimativas para uma taxa de recombinação constante ao longo do genoma.

2.2.4.5 O método de Li & Stephens (2003)

Li & Stephens (2003) apresentaram um modelo que supera várias limitações dos modelos anteriores. Segundo esses autores, o novo método: i) relaciona os padrões do desiquilíbrio de ligação diretamente ao processo de recombinação; ii) considera todos os locos simultaneamente, ao invés de considerá-los aos pares; iii) evita a suposição de que o desequilíbrio de ligação tem, necessariamente, uma estrutura em blocos; e iv) é computacionalmente praticável até mesmo para cromossomos completos. Os autores ainda incluíram no modelo a possibilidade de estimar taxas de recombinação variáveis ao longo do genoma e de inferir sobre a presença e intensidade de *hotspots* de recombinação.

Como o propósito da presente pesquisa foi incluir a caracterização da variação da taxa de recombinação ao longo do genoma de *A. thaliana*, a seguir é apresentada uma descrição detalhada do modelo proposto por Li & Stephens (2003).

De acordo com a categorização de Stephens (2007), o modelo estatístico apresentado por Li & Stephens (2003) é um modelo que faz uso do conjunto completo de dados genéticos, sendo baseado no modelo de coalescência e faz uso dos algoritmos MCMC, ou seja, é um método baseado em aproximações às verossimilhanças, conforme

classificação de Stumpf & McVean (2003).

O modelo relaciona a distribuição de uma amostra de haplótipos com a taxa de recombinação subjacente, por meio da seguinte igualdade:

$$\Pr(h_1, \ldots, h_n \mid \rho) = \Pr(h_1 \mid \rho)\, \Pr(h_2 \mid h_1; \rho) \ldots \Pr(h_n \mid h_1, \ldots, h_{n-1}; \rho) \quad (1)$$

Em que:

$h_1, \ldots, h_n$: representa os n haplótipos amostrados;

ρ : parâmetro da taxa de recombinação, que pode assumir a forma de um vetor de parâmetros, caso se permita que a taxa de recombinação varie ao longo do genoma.

Essa equação expressa uma distribuição de probabilidade desconhecida do lado esquerdo, como função do produto de distribuições condicionais do lado direito. π será a notação utilizada para representar estas distribuições condicionais. Como as distribuições condicionais podem ser computacionalmente muito intensivas, elas podem ser substituídas por aproximações capazes de serem processadas.

Li & Stephens (2003) usam a estratégia de substituir as distribuições condicionais do lado direito da equação (1) por uma aproximação denotada $\hat{\pi}$, de modo a obter uma aproximação para a distribuição dos haplótipos h dada uma taxa de recombinação ρ:

$$\Pr(h_1, \ldots, h_n \mid \rho) \approx \hat{\pi}(h_1 \mid \rho)\ \hat{\pi}(h_2 \mid h_1; \rho) \ldots \hat{\pi}(h_n \mid h_1, \ldots, h_{n-1}; \rho) \quad (2)$$

A descrição apresentada em (2) é denominada de modelo dos "produtos das aproximações às condicionais" (PAC). Sua respectiva verossimilhança é denominada de verossimilhança dos "produtos das aproximações às condicionais", isto é, verossimilhança-PAC, cuja notação é L_{PAC}.

$$L_{PAC}(\rho) = \hat{\pi}(h_1 \mid \rho)\ \hat{\pi}(h_2 \mid h_1; \rho) \ldots \hat{\pi}(h_n \mid h_1, \ldots, h_{n-1}; \rho) \quad (3)$$

De modo análogo, o valor de ρ que maximiza L_{PAC} é denominado de estimativa de máxima verossimilhança de PAC para ρ, cuja notação é $\hat{\rho}_{PAC}$. Então, a aplicabilidade do modelo (3) dependerá da escolha de uma aproximação $\hat{\pi}$ adequada para representar, da melhor forma possível, a distribuição condicional π. Essa aproximação deve ser elaborada de modo a responder ao seguinte questionamento: se, para um loco qualquer, numa amostra aleatória de k cromossomos de uma população, sejam observados os haplótipos h_1, . . . , h_k, qual é a distribuição condicional para o haplótipo do próximo cromossomo a ser amostrado, $\Pr(h_{k+1} \mid h_1, \ldots, h_k)$?

Li & Stephens (2003) descrevem cinco formas distintas de aproximações para a distribuição condicional π. Estas foram elaboradas por Ewens (1972), Stephens & Donnelly (2000) e Fearnhead & Donnelly (2001) e duas outras formas elaboradas por eles mesmos. Cada uma das cinco formas apresenta uma resposta para a questão, sob diferentes suposições associadas ao modelo genético subjacente aos locos em estudo.

No método de Ewens (1972) considera-se que a população é panmítica e evolui com tamanho constante (N), os locos são neutros e sofrem uma taxa de mutação (μ) por geração, assumindo-se o modelo mutacional de alelos infinitos. Sob todas essas condições, se $\theta = 4N\mu$, então o $(k+1)^{ésimo}$ haplótipo será uma cópia exata dos k primeiros haplótipos escolhidos aleatoriamente com a probabilidade de $k/(k+\theta)$; senão o $(k+1)^{ésimo}$ haplótipo será um haplótipo novo. Apesar de que as suposições impostas ao modelo genético raramente ocorrem na prática, o método de Ewens (1972) tem as seguintes propriedades:

i. O próximo haplótipo a ser amostrado tem maiores chances de ser um haplótipo já observado várias vezes do que um haplótipo observado com menor frequência;

ii. A probabilidade de se obter um haplótipo novo diminui à medida que k aumenta;

iii. A probabilidade de se obter um haplótipo novo aumenta à medida que θ aumenta.

Porém, para dados de sequência e, em particular, para dados de marcadores SNP, o método de Ewens (1972) falha em captar duas outras propriedades:

iv. Se o próximo haplótipo a ser amostrado não for exatamente igual a um haplótipo já observado, ele tenderá a diferir dos haplótipos já observados em apenas um pequeno número de mutações sobre o haplótipo observado, ao invés de ser completamente diferente de todos os haplótipos já observados;

v. Pelo efeito da recombinação, o próximo haplótipo a ser amostrado tenderá a ser semelhante aos haplótipos já observados ao longo de regiões contínuas do genoma, sendo que o comprimento físico médio dessas regiões será maior em áreas do genoma onde a taxa de recombinação é relativamente mais baixa.

A forma de π proposta por Stephens & Donnelly (2000) é dotada das propriedades *i*, *ii*, *iii* e *iv* anteriormente apresentadas. Nesta forma, o próximo haplótipo a ser amostrado difere, em M mutações, de um haplótipo existente escolhido ao acaso. M

obedece uma distribuição geométrica com $Pr(M=0) = k/(k+\theta)$. Dessa forma, o próximo haplótipo a ser amostrado será uma cópia (possivelmente imperfeita) de um haplótipo existente escolhido ao acaso.

Fearnhead & Donnelly (2001) elaboraram uma extensão à forma de π de Stephens & Donnelly (2000), para captar a propriedade *v*. Nesta forma, o $(k+1)^{ésimo}$ haplótipo a ser amostrado é composto de um mosaico de partes dos primeiros *k* haplótipos existentes, sendo que o tamanho dos fragmentos desse mosaico será menor em regiões onde a taxa de recombinação é relativamente mais alta.

As duas novas formas de π (π_A e π_B) apresentadas por Li & Stephens (2003) também captam as cinco propriedades. Uma representação gráfica desta proposta está ilustrada na Figura 13.

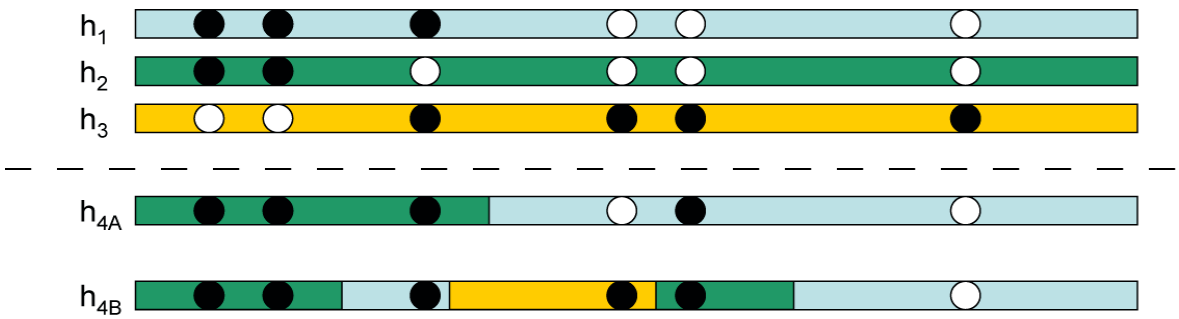


Figura 13. Demonstração gráfica de como $\pi_A(h_{k+1}|h_1, \ldots, h_k)$ constrói um haplótipo h_{k+1} simulando um mosaico imperfeito dados os haplótipos $h_1, \ldots, h_k$. Supondo k=3 a ilustração mostra duas possibilidades para h_4 (h_{4A} e h_{4B}), dado h_1, h_2 e h_3. Tanto h_{4A} quanto h_{4B} podem ser imaginados como tendo sido criados a partir de cópias (às vezes não perfeitas) de partes de h_1, h_2 e h_3. As cores mostram qual haplótipo foi copiado em cada posição ao longo do cromossomo. É pressuposto que o processo de cópia é um processo markoviano ao longo do cromossomo, com saltos (isto é, mudanças nas cores), que ocorrem a uma taxa ρ/k por unidade de distância física. Portanto, as mudanças mais freqüentes de cor no haplótipo h_{4B} sugerem maior valor de ρ quando comparado com mudanças menos freqüentes que se observa no h_{4A}. Para valores extremamente altos de ρ, os locos se tornam independentes, como seria de se esperar. Cada coluna de círculo representa um loco SNP, com as cores branca e preta representando os dois alelos. A natureza imperfeita do processo de cópia é exemplificada no quinto loco, no qual os haplótipos h_{4A} e h_{4B} têm o alelo preto, apesar de eles terem copiado os haplótipos h_1 e h_2, respectivamente, que têm o alelo branco. Na prática, as cores não são observadas, exigindo que o cálculo da probabilidade de se observar um haplótipo h_4 em particular seja feito com base na soma de todas as possibilidades de combinações de cores. A verossimilhaça de um dado mosaico é uma função do número de eventos de recombinação e de mutação ao longo do cromossomo. Fonte: Li & Stephens (2003).

A seguir apresenta-se a descrição formal para a aproximação à distribuição condicional π_A, formulada por Li & Stephens (2003), extraída desta mesma fonte:

Considere $h_1, \ldots, h_n$ como sendo uma amostra de n haplótipos, genotipados em S locos bialélicos (SNPs). Em geral, $h_1, \ldots, h_n$ podem ter origem numa amostra de n indivíduos haplóides ou de $n/2$ indivíduos diplóides. Assume-se que a distribuição do primeiro haplótipo é independente de ρ; isto é, todos os 2^S possíveis haplótipos tem mesma probabilidade de ocorrência, ou seja $\pi_A(h_1) = 1/2^S$.

Considere agora a distribuição condicional de h_{k+1}, dado $h_1, \ldots, h_k$, para $k \geq 1$. Observando-se a Figura 13, verifica-se que h_{k+1} é um mosaico imperfeito de $h_1, \ldots, h_k$. Para $k \geq 1$, na posição de cada SNP, o haplótipo h_{k+1} se constituirá de uma cópia (possivelmente imperfeita) de um dos $h_1, \ldots, h_k$ haplótipos, naquela referida posição.

Seja X_j a identificação de qual dos $h_1, \ldots, h_k$ haplótipos o haplótipo h_{k+1} se copiou na posição j [$X_j \in \{1, 2, \ldots, k\}$]. Por exemplo, para o haplótipo h_{4A}, na Figura 13, ($X_1, X_2, X_3, X_4, X_5, X_6$) = (2, 2, 2, 1, 1, 1); já para o haplótipo h_{4B}, ($X_1, X_2, X_3, X_4, X_5, X_6$) = (2, 2, 1, 3, 2, 1). Para simular os efeitos da recombinação modelou-se X_j como uma cadeia de Markov em $\{1, \ldots, k\}$, com

$\Pr(X_1 = x) = 1 / k \quad [x \in \{1, \ldots, k\}]$, e

$$\Pr(X_{j+1} = x' \mid X_j = x) = \begin{cases} \exp\left(\frac{-\rho_j d_j}{k}\right) + \left(1 - \exp\left(\frac{-\rho_j d_j}{k}\right)\right)\left(\frac{1}{k}\right) & \text{se x' = x;} \\ \left(1 - \exp\left(\frac{-\rho_j d_j}{k}\right)\right)\left(\frac{1}{k}\right) & \text{, em outros casos.} \end{cases}$$

em que:

d_j : distância física entre as marcas j e $j+1$ (assume-se conhecida);

$\rho_j = 4Nc_j$;

N : tamanho efetivo da população;

c_j : taxa média de *crossover* por unidade de distância física, por meiose, entre as posições j e $j+1$ (de modo que $c_j d_j$ é a distância genética entre as posições j e $j+1$).

Para a condição em que as posições j e $j+1$ estejam a uma pequena distância genética entre si (isto é, $c_j d_j$ é bastante pequena), essa matriz de transição fará com que ambas as posições tenham grandes possibilidades de copiarem o mesmo cromossomo (isto é, $X_{j+1} = X_j$).

Explicitando-se alguns casos especiais:

a) taxa de recombinação constante: $c_j = \overline{c}$, para todo e qualquer *j*;

b) modelo com apenas um *hotspot*; $c_j = \overline{c}$ se as posições *j* e *j+1* estão ambas fora do *hotspot*, e $c_j = \lambda\overline{c}$ se as posições *j* e *j+1* estão ambas dentro do *hotspot*;

c) taxa de recombinação variável ao longo de uma região: $c_j = \lambda\overline{c}$.

Para modelar-se os efeitos da mutação, o processo de cópia pode ser, às vezes, imperfeito. Com a probabilidade $k/\left(k+\tilde{\theta}\right)$ a cópia é perfeita, e com a probabilidade $\tilde{\theta}/\left(k+\tilde{\theta}\right)$, uma "*mutação*" é aplicada ao haplótipo copiado. Seja $h_{i,j}$ a representação do alelo (0 ou 1) na posição *j* no haplótipo *i*, então, considerando-se o processo de cópia $X_1, \ldots, X_S$, os alelos $h_{k+1,1}, h_{k+1,2}, \ldots, h_{k+1,S}$ são independentes, com

$$\Pr\left(h_{k+1,j} = a \mid X_j = x, h_1, \ldots, h_k\right) = \begin{cases} k/(k+\tilde{\theta}) + (1/2)\left(\tilde{\theta}/\left(k+\tilde{\theta}\right)\right), & h_{x,j} = a; \\ (1/2)\left(\tilde{\theta}/\left(k+\tilde{\theta}\right)\right), & h_{x,j} \neq a. \end{cases} \quad (5)$$

A constante ½ aparece em ambos os casos, e à medida que $\tilde{\theta} \to \infty$, ambos os alelos passam a ter a mesma probabilidade de ocorrência. Então, o valor de $\tilde{\theta}$ foi fixado em:

$$\tilde{\theta} = \left(\sum_{m=1}^{n-1} \frac{1}{m}\right)^{-1} \quad (6)$$

Em que: *n* é o número total de haplótipos amostrados.

Os cálculos matemáticos de $\pi_A(h_{k+1}|h_1, \ldots, h_k)$ requer uma soma de todos os possíveis valores de X_j, a qual pode ser realizada de modo eficiente usando o módulo *forward* do algoritmo *forward-backward* da família de modelos de Markov ocultos, Rabiner[2] (1989) , citado por Li & Stephens (2003).

Considere que $h_{k+1,\leq j}$ representa o tipo das primeiras *j* regiões do haplótipo h_{k+1}, e que $\alpha_j(x) = \Pr(h_{k+1,\leq j}, X_j = x)$. Então, $\alpha_1(x)$ pode ser calculada diretamente para $x = 1, \ldots, k$. Já $\alpha_2(x), \ldots, \alpha_S(x)$ podem calculadas recursivamente usando a expressão:

---

[2] RABINER, L. R. A tutorial on hiden Markov models and selected applications in speech recognition. **Proceedings of the IEEE**, v. 77, n. 2, p. 257-286, feb. 1989.

$$\alpha_{j+1}(x) = \gamma_{j+1}(x)\sum_{x'=1}^{k}\alpha_j(x')\Pr(X_{j+1} = x \mid X_j = x') \qquad (7)$$

$$\alpha_{j+1}(x) = \gamma_{j+1}(x)\left(p_j\alpha_j(x) + \left(1 - p_j\right)\frac{1}{k}\sum_{x'=1}^{k}\alpha_j(x')\right) \qquad (8)$$

Em que:

$\gamma_{j+1}(x) = \Pr\left(h_{k+1,j+1} = a \mid X_{j+1} = x,\ h_1,\ \ldots, h_k\right)$ é dada na expressão (5), e

$p_j = \exp(-\rho_j d_j / k)$.

Então, o valor de $\pi_A(h_{k+1}|h_1, \ldots, h_k)$ pode ser calculado usando-se a seguinte expressão:

$$\pi_A\left(h_{k+1} \mid h_1, \ldots, h_k\right) = \sum_{x=1}^{k}\alpha_s(x). \qquad (9)$$

O segundo termo dentro do parêntesis na expressão (8) não depende de x e, portanto, basta ser calculado uma vez para cada j. Daí, infere-se que a complexidade computacional no cálculo de $\pi_A(h_{k+1}|h_1, \ldots, h_k)$ cresce linearmente com o número de SNP e linearmente com k. Consequentemente, a complexidade do cálculo de $L_{PAC\text{-}A}$ cresce linearmente com o número de SNPs e de forma quadrática com o número de cromossomos da amostra.

Convém ressaltar que nenhuma das formas de π apresentadas por Stephens & Donnelly (2000), Fearnhead & Donnelly (2001), Hudson (2001), McVean et al. (2002) e Li & Stephens (2003) representa, fielmente, a real distribuição condicional correspondente a um modelo populacional sobre o qual estejam atuando os efeitos demográficos e das forças evolutivas mais preponderantes. Todas são aproximações se baseiam em uma ou outra suposição. Até o presente momento ainda não existe expressão para π, que capte as cinco propriedades, e que, concomitantemente, modele explicitamente os efeitos dos vários fatores evolutivos e suas interações.

Um comentário, especificamente sobre as formas de π apresentadas por Stephens & Donnelly (2000) e por Li & Stephens (2003), é que as estimativas de $\hat{\rho}_{PAC}$ dependem da ordem em que os haplótipos são considerados. Stephens & Donnelly (2000) sugerem que esse problema poderia ser contornado fazendo-se a média de $\hat{\rho}_{PAC}$ usando todas as possíveis ordens de haplótipos. Isto acarretaria no somatório de $n!$ termos, algo impraticável mesmo para pequenos valores de n. A alternativa utilizada por esses autores

foi fazer uma média sobre um número menor de ordens, escolhidas aleatoriamente. No *software* desenvolvido por Li & Stephens (2003), $\hat{\rho}_{PAC}$ é estimada com base na média de vinte ordens aleatoriamente escolhidas.

A comparação da eficiência da forma de π apresentada por Li & Stephens (2003) com as anteriores foi feita usando-se dados simulados. Foram gerados dados sob várias combinações de *n* (número de haplótipos), *S* (número de marcas) e ρ (taxa de recombinação). Para cada conjunto de dados os autores calcularam $\hat{\rho}_{PAC-A}$ maximizando numericamente a verossimilhança-PAC e comparando a estimativa obtida com o valor verdadeiro de ρ usado para gerar os dados. As comparações entre as estimativas $\hat{\rho}_{PAC-A}$ e o parâmetro ρ foram feitas em uma escala relativa, ao invés de usar uma escala absoluta. O erro relativo de uma estimativa de $\hat{\rho}$ para o verdadeiro valor de ρ foi calculado usando:

$$Erro(\rho, \hat{\rho}) = \log_{10}(\frac{\hat{\rho}}{\rho}) \qquad (10)$$

Essa expressão fornece, por exemplo, um erro com valor zero, se $\hat{\rho} = \rho$; um erro com valor 1 se $\hat{\rho}$ superestimar ρ por um fator de dez vezes para mais; e um erro com valor -1 se $\hat{\rho}$ subestimar ρ por um fator de dez vezes para menos.

Após intensas simulações, Li & Stephens (2003) verificaram que para algumas combinações de *n*, *S* e ρ os valores de $Erro(\rho, \hat{\rho})$ mostraram uma distribuição normal e centrada em zero. Já para outras combinações *n*, *S* e ρ os valores de $Erro(\rho, \hat{\rho})$ continuavam obedecendo uma distribuição próxima da normal mas com um viés, ou seja centrada em algum valor diferente de zero, como ilustrado na Figura 14.

Apesar do viés de $\hat{\rho}_{PAC-A}$ depender das três variáveis *n*, *S* e ρ, este é mais fortemente dependente do espaçamento médio entre os pontos; isto é, para um dado par *n* e *S* foi observado uma relação linear entre o viés e o logaritmo do espaçamento médio entre pontos. Como a inclinação da relação linear é negativa, há uma tendência de $\hat{\rho}_{PAC-A}$ superestimar ρ quando as marcas estão muito próximas entre si, e de subestimar ρ quando as marcas estão muito distantes entre si.

Com base nessas observações, Li & Stephens (2003) modificaram π_A, criando π_B, uma nova versão de π_A contendo uma correção empírica para seu viés. Para corrigir os viéses com o estimador $\hat{\rho}_{PAC-A}$, eles modificaram a matriz de transição da expressão (4),

substituindo ρ_j por $\delta_j\rho_j$, em que $\delta_j = \exp(a + b \times \log_{10} \rho_j)$. O intercepto a e o coeficiente angular b são interpolados com base no número de haplótipos n e número de pontos segregantes S nos dados, usando modelos particulares de interpolação baseados em *splines* cúbicas, Ueberhuber[3] (1997), citado por Li & Stephens (2003).

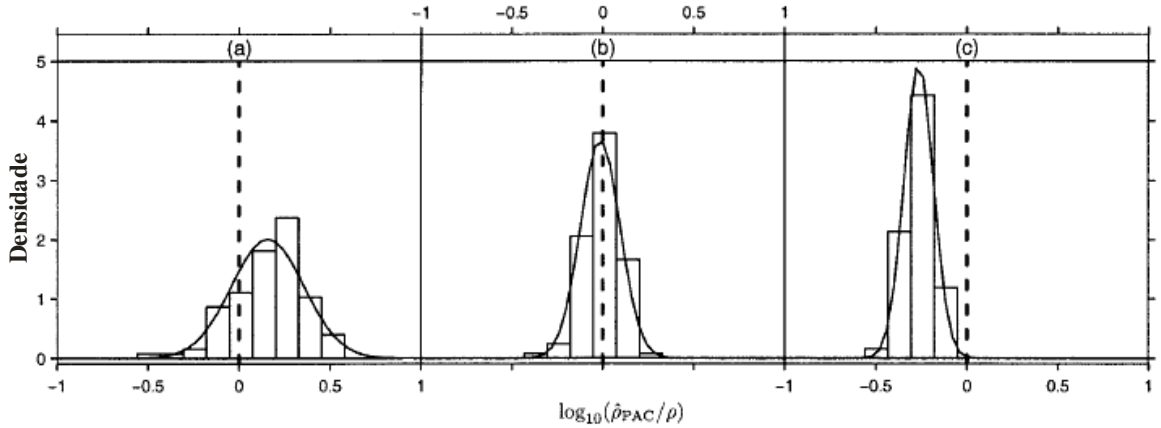


Figura 14. Histogramas dos erros $Erro(\rho,\hat{\rho}_{PAC-A})=\log_{10}(\hat{\rho}_{PAC-A}/\rho)$, baseados em cem conjuntos de dados simulados, a partir do modelo de coalescência padrão usando $n = 50$ haplótipos e $S = 50$ pontos segregantes. Os valores de ρ são: (a) $\rho = 5$, (b) $\rho = 25$, (c) $\rho = 500$. As curvas sobrepostas às barras representam a densidade da função normal com a mesma média e desvio padrão dos cem valores usados para compor o histograma. Os resultados são baseados em médias das verossimilhanças de dez ordens aleatórias de haplótipos. Fonte: Li & Stephens (2003).

Ao considerarem uma taxa de recombinação variável ao longo do genoma, os autores Li & Stephens (2003) formalizaram dois modelos para explorar essa variação em escalas locais. Os dois modelos são baseados em um modelo geral que considera a ocorrência de *crossover* em uma meiose como um processo de Poison, com taxa *c(x)* numa posição *x*.

O primeiro modelo considera um único *hotspot* de recombinação:

$$c(x) = \begin{cases} \lambda\bar{c} & a \le x \le b\ , \\ \bar{c} & \text{para os demais casos.} \end{cases} \qquad (11)$$

Em que:

$\bar{c}$: representa a taxa média de fundo de ocorrência de *crossover*;

a e *b* : representam as posições de início e término da região do *hotspot* ;

λ (>1) : quantifica a intensidade da atividade do *hotspot* de recombinação.

A PAC-verossimilhança para esse modelo é uma função de quatro parâmetros:

[3] UEBERHUBER, C. W. **Numeric computation: methods, software and analysis**. Vol. I. Berlin: Springer-Verlag, 1997. 474 p.

a, b, λ e $\overline{\rho} = 4N\overline{c}$. A evidência para a detecção da presença de um *hotspot* de recombinação pode ser obtida pela razão de verossimilhança entre a hipótese de nulidade, correspondente a não existência de *hotspot* (H_0: $\lambda = 1$), e a hipótese alternativa (H_1: $\lambda > 1$). Considerando que $\overline{\rho}_0$ representa o valor de $\overline{\rho}$ que maximiza $L_{PAC\text{-}B}$, sob a hipótese H_0, e que $\overline{\rho}_1$ e λ_1 representam os valores de $\overline{\rho}$ e λ que maximizam $L_{PAC\text{-}B}$, sob a hipótese H_1, então:

$$LLR = \log_e \left[L_{PAC-B}(\overline{\rho}_1, \lambda_1) / L_{PAC-B}(\overline{\rho}_0, \lambda = 1) \right], \quad (12)$$

e, assim, altos valores de *LLR* representam evidências da existência de um *hotspot*. Segundo Li & Stephens (2003), o dobro do valor de *LLR* tem distribuição de χ^2 com 1 grau de liberdade e, sob esta suposição, rejeitar H_0 para valores de *LLR* maiores que 1,92 resultaria num teste de hipótese com, no máximo, 5% de chance de cometer o erro tipo I.

O segundo modelo considera que se x é uma posição entre as marcas j e $j+1$, então:

$$c(x) = \lambda_j \overline{c} \qquad (13)$$

Em que:

λ_j : é um multiplicador que controla o quanto a taxa de *crossover* entre as marcas j e $j+1$ se desvia da taxa média de ocorrência de *crossover*.

A verossimilhança-PAC para este modelo é uma função dos parâmetros λ_1 , . . . , λ_{S-1} (em que S é o número de SNP) e $\overline{\rho} = 4N\overline{c}$.

Li & Stephens (2003) relatam que a tentativa de obtenção das estimativas por meio da verossimilhança-PAC para os parâmetros gera dois problemas. O primeiro refere-se ao fato de que as estimativas de máxima verossimilhança não são únicas, fato designado por *não-identificabilidade* dos parâmetros. O segundo refere-se ao caso em que as curvas de verossimilhança, para alguns valores de λj, serão sempre achatadas, sem picos, resultando em estimativas de λ_j ou muito próximas de zero ou tendendo ao infinito. Se as curvas de verossimilhança para determinados λ_j são bastante achatadas, isso indica que existe pouca informação sobre a taxa de recombinação naquele intervalo entre marcas. Nestes casos seria de bom senso supor que a taxa de recombinação no intervalo assumiria valores próximos da taxa média de recombinação (isto é, $\lambda_j \approx 1$) e não valores infinitamente pequenos ou grandes.

Para resolver esses dois tipos de problemas os autores propuseram que fosse adotada uma distribuição *a priori* para os λ_j: os valores de λ_j são independentes e

identicamente distribuídos com $\log_{10}(\lambda_j) \sim N [0, (½)^2]$. Essa distribuição *a priori* foi escolhida com base em uma justificativa mais pragmática de permitir que estimativas da taxa de recombinação obtidas num intervalo possam ter desvios de até dez vezes, para mais ou para menos, em relação à taxa de recombinação média.

2.2.4.6 O método de McVean et al. (2004)

McVean et al. (2004) desenvolveram um método baseado em aproximações à verossimilhança, a partir de uma extensão ao estimador de verossimilhança composta de Hudson (2001). Eles modelaram a taxa de recombinação, permitindo que esta varie de um par de SNP para outro. Os autores adotaram uma abordagem bayesiana, na qual o uso de uma distribuição *a priori* permite a suavização das estimativas da taxa de recombinação dentro do alcance de alguns intervalos mais próximos entre si. O grau de suavização é proporcional a um fator de penalização para mudanças na taxa de recombinação. Esse método permite a estimação da taxa de recombinação em várias escalas, desde escalas no nível de kilobases, até escalas dos atuais mapas genéticos. O método é aplicado tanto a dados de genótipos, quanto de haplótipos. É bastante robusto aos problemas de genotipagem de SNP. Permite optar pelo uso do modelo de mutações recorrentes, lida bem com a falta de dados e é computacionalmente viável sobre conjuntos de dados mais densos.

McVean et al. (2004) informaram que o uso de um valor zero para o fator de penalização propiciou uma superestimação para as taxas de recombinação. Aumentando-se este fator corrige-se o viés do estimador e evita-se a detecção de variações espúrias na taxa de recombinação. Os autores sugerem que vinte é o valor mais adequado para o fator de penalização. Uma das desvantagens deste método é essa determinação empírica do valor mais apropriado para o fator de penalização ou suavização.

2.2.4.7 O método de Fearnhead & Smith (2005)

O método proposto por Fearnhead & Smith (2005) é baseado no método de aproximação às verossimilhanças de Fearnhead & Donnelly (2002). O método divide uma região em sub-regiões de, no máximo, seis SNP consecutivos. Para cada sub-região: a) é realizada a simulação de um conjunto de 100.000 possíveis genealogias para os dados; b) é calculada uma aproximação à superfície de verossimilhança parcial para a taxa de

recombinação de cada uma das 100.000 genealogias; c) é calculada a combinação de todas verossimilhanças parciais em uma verossimilhança composta para a sub-região, usando-se uma média ponderada com pesos adequados para cada genealogia; d) é calculado o produtório das verossimilhanças de todas as sub-regiões que comporá a curva de verossimilhança para a região completa. O modelo de *hotspot* de recombinação presente neste método considera apenas os eventos de recombinação homóloga, sem modelar os eventos de conversão gênica.

Segundo os próprios autores deste método, aqueles de Li & Stephens (2003) e de McVean et al. (2004) são mais indicados para os casos em que se tem uma menor densidade de SNP.

2.2.4.8 O método de Fearnhead (2006)

O método proposto por Fearnhead (2006) também é baseado em aproximações à verossimilhança de Fearnhead & Donnelly (2002), sendo bastante semelhante ao método de Fearnhead & Smith (2005). Nesse método, uma região é divida em sub-regiões e, para cada sub-região, é calculada uma razão de verossimilhanças para testar a evidência de existência de um *hotspot* de recombinação. O conjunto de todos os valores das razões de verossimilhança pode ser usado para mostrar, visualmente, uma medida de evidência para a presença de *hotspots* de recombinação ao longo do cromossomo.

Esse método é computacionalmente mais rápido por duas razões: *i*) a qualquer momento os cálculos das razões de verossimilhanças podem ser abortados, caso se torne óbvio que há poucas evidências da presença de um *hotspot*; *ii*) são obtidas estimativas mais precisas para a taxa de recombinação de fundo, usando-se o estimador $\hat{\rho}_{PAC-B}$ construído por Li & Stephens (2003). Também é um método recomendado para regiões com baixa densidade de SNP.

2.2.4.9 O método de Auton & McVean (2007)

Esse método consiste num aprimoramento do método de McVean et al. (2004), substituindo as distribuições *a priori* para a taxa de recombinação de fundo e para os *hotspots*. Auton & McVean (2007) consideraram *a priori* usada por McVean et al. (2004), para a variação da taxa de recombinação, como um modelo pouco representativo dos

verdadeiros níveis de variação observados na taxa de recombinação.

Eles descrevem uma *priori* mais sofisticada e usam o esquema *rjMCMC* (*reversible jump Markov Chain Monte Carlo*) para obter as estimativas da taxa de recombinação. Na função densidade de probabilidade usada como *priori*, no método de McVean et al. (2004), a taxa de recombinação era modelada como sendo constante dentro de um intervalo, entre um SNP e outro, variando seus valores apenas quando ocorria uma mudança de intervalo. A nova *priori* para a taxa de recombinação permite a mudança nas estimativas da taxa de recombinação em qualquer ponto dentro ou no limite dos intervalos entre SNP. Se anteriormente as posições de mudanças da taxa de recombinação estavam amarradas às posições dos SNP, agora as posições de mudança têm uma distribuição probabilística ao longo do cromossomo. A Figura 15 ilustra a diferença nas estimativas da taxa de recombinação realizadas pelos dois métodos.

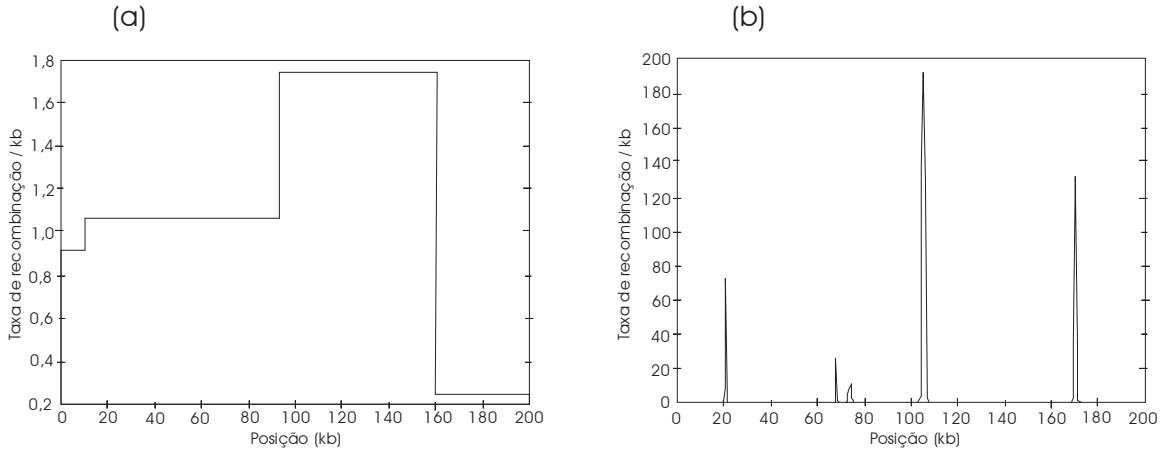


Figura 15. Ilustração da *priori* usada nos métodos: (a) de McVean et al. (2004), e (b) de Auton & McVean (2007). Observa-se a diferença de escala nos eixos Y dos dois gráficos. Fonte: Auton (2007).

Auton & McVean (2007) também incluíram a descrição de um novo modelo para os *hotspots* de recombinação. Nesse método os *hotspots* são modelados como picos agudos na taxa de recombinação, lembrando os gráficos de funções exponenciais duplas do tipo $f(x) = a^{b^x}$. O número de *hotspots* e suas propriedades (posição, largura e intensidade) são determinadas como parte de um esquema *rjMCMC*. Uma importante contribuição de Auton & McVean (2007) é modelagem da morfologia de um *hotspot* (Figura 16), que passa a ser caracterizado por sua posição, por sua extensão e por sua intensidade ou grau de atividade. Nesse método, a *priori* usada para descrever a distribuição dos *hotspots* os consideram como uniformemente espalhados ao longo da

região.

Um *hotspot* é representado por uma função exponencial dupla com escala μ, truncada em ambas as caudas. A largura de um *hotspot* é definida pela região sob a curva da função exponencial dupla que contém 95% de sua massa. A intensidade da atividade de um *hotspot* é representada por um parâmetro λ. O corte nas caudas da curva exponencial dupla é feito sempre que se encontra um ponto de mudança, ou, caso não exista ponto de mudança nas proximidades, por meio de uma distância arbitrária *m* a partir do pico do *hotspot*.

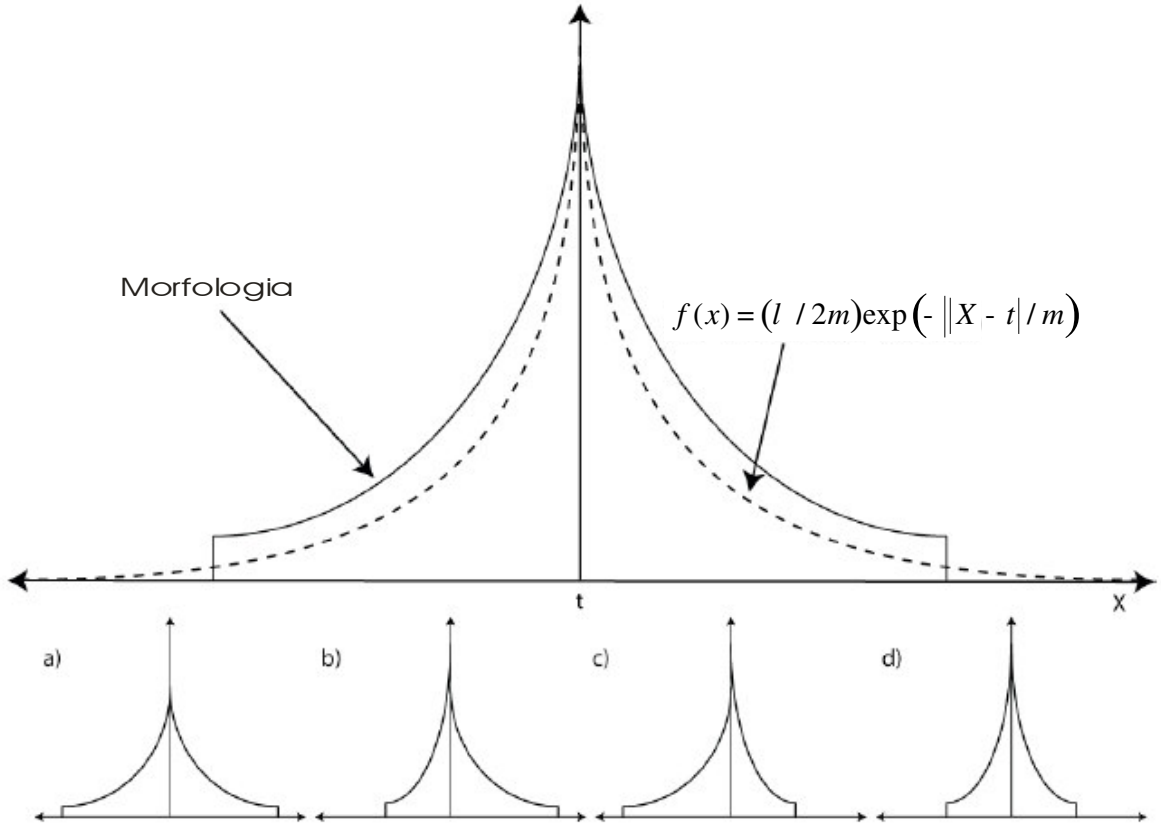


Figura 16. Morfologia de um *hotspot* de recombinação. A massa das partes truncadas nas caudas é adicionada ao corpo principal da distribuição. Quatro cenários demonstram como a morfologia se altera em função dos pontos de mudança da taxa de recombinação: a) nenhum ponto de mudança da taxa de recombinação nas proximidades do *hotspot*, fazendo com que sua largura se extenda até a máxima permitida (*m*); b) um ponto de mudança da taxa de recombinação está próximo e à esquerda do *hotspot*; c) um ponto de mudança da taxa de recombinação está próximo e à direita do *hotspot*; d) dois pontos de mudança da taxa de recombinação estão próximos ao *hotspot*, um à esquerda e outro à direita. A área sob a curva é a mesma em todos os *hotspots*. Fonte: Auton (2007).

2.2.4.10 O método de Wang & Rannala (2008)

A proposta apresentada por Wang & Rannala (2008) consiste num método baseado em funções de verossimilhanças completas, sem aproximações, dentro de um contexto de estatística bayesiana. Apesar de fazer uso das verossimilhanças puras completas, os autores prometem que o método tem uma demanda computacional aceitável.

A principal evolução em relação ao método de Auton & McVean (2007) é que, para regiões de tamanho moderado (exemplo, $\leq$ 100 SNP), a probabilidade *a posteriori* das taxas de recombinação é obtida a partir de uma função de verossimilhança completa, evitando-se as aproximações seguidas de produtórios para gerar a verossimilhança composta daquela região. Se a região é mais ampla, contém um maior número de SNP, então, ela é dividida em *n* partes. Os autores sugerem que *n* fique sempre entre 20 e 50. A verossimilhança composta é, então, calculada condicionada às genealogias de cada sub-região e, em seguida, os parâmetros são avaliados de forma conjunta usando-se todas as sub-regiões.

No modelo de Wang & Rannala (2008) é assumido que a distribuição dos *hotspots* de recombinação, ao longo de um cromossomo, obedece um processo Markoviano. Um *hotspot* surge com uma taxa instantânea λ_1 e desaparece com uma taxa instantânea λ_2. A distância percorrida até que surja um *hotspot* tem distribuição exponencial com parâmetro λ_1, e a distância percorrida até o desaparecimento do *hotspot* em questão também tem distribuição exponencial, mas com parâmetro λ_2. Os valores de $1/\lambda_1$ e $1/\lambda_2$ representam, respectivamente, a distância média entre *hotspots* e a duração média dos *hotspots*. Três variáveis são associadas a cada *hotspot*: X_1, que representa a posição de início; X_2, que representa a posição de término; e Z, que representa a intensidade do *hotspot*. Os autores assumem que Z tem distribuição log-normal com parâmetros μ_z e σ_z.

No método de Wang & Rannala (2008) assume-se que a taxa de recombinação entre SNP é independentemente distribuída segundo uma distribuição gama (distribuição-Γ), com parâmetro de forma a_{ρ^*} e parâmetro de escala S_{ρ^*}. Durante as análises, S_{ρ^*} é fixado e a_{ρ^*} é estimado por um algoritmo *MCMC*.

2.2.4.11 O método de Lefebvre & Labuda (2008)

A metodologia ou o algoritmo apresentado por Lefebvre & Labuda (2008) resolve o problema da estimação da taxa de recombinação populacional usando dados genéticos de populações de uma maneira bastante satisfatória. A denominação dada à metodologia foi "Metodologia das Recombinações Informativas", implementada no *software* InfRec, disponível gratuitamente mediante solicitação aos autores.

O método é intuitivo e combina partes de métodos desenvolvidos anteriormente. Para estimar haplótipos a partir de dados genotípicos é incorporada a metodologia desenvolvida por Stephens et al. (2001), cujo algoritmo foi implementado no *software* PHASE. Para estimar o número mínimo de recombinações numa amostra é utilizada a metodologia elaborada por Myers & Griffiths (2003) e para estimar as frações de recombinações informativas (*FIR – Fraction of Informativa Recombinations*) e de recombinações novas (*FNR – Fraction of Novel recombinations*) é incorporada a metodologia que foi desenvolvida por Zietkiewicz et al. (2003).

O *software* InfRec é fácil de ser utilizado, tem uma interface gráfica, não impõe limite para o tamanho das amostras e tem uma demanda computacional relativamente menor do que a dos programas de computador que implementam algoritmos *MCMC*, para estimar funções de verosimilhança completas. O método fornece uma relação transparente entre os dados observados e as estimativas resultantes de sua aplicação. Quando se compara as estimativas da taxa de recombinação geradas pelo *software* InfRec com as geradas por outros modelos e programas de computador, o InfRec pode ser considerado bastante satisfatório.

O desempenho do InfRec em captar *hotspots* de recombinação foi testado usando-se dados simulados pelo modelo de coalescência implementado no programa msHOT, de autoria de Hellenthal & Stephens (2007), e quatro outros conjuntos de dados de regiões do genoma humano, cujos *hotspots* de recombinação já são conhecidos e bem caracterizados. Os parâmetros de avaliação foram: a taxa de falsas detecções (*FDR – False Discovery Ratio*) por milhões de pares de bases, e o poder de detecção dos *hotspots*, calculado com base na proporção dos *hotspots* simulados que foram encontrados pelas estimativas geradas pelo InfRec. As estimativas das taxas de recombinação ao longo das regiões genômicas analisadas estão em excelente concordância com estimativas publicadas anteriormente. Porém, conforme ressaltam os próprios autores desse método (Lefebvre &

Labuda, 2008), as estimativas geradas pelo InfRec podem ser consideradas mais conservadoras, pois se baseiam no número mínimo de recombinações estimadas pelo *software* RecMin. Quanto ao poder de detecção dos *hotspots*, o *software* InfRec obteve taxas entre 79% e 89% do número total de *hotspots* existentes em todos os conjuntos de dados analisados. A taxa de falsas descobertas por milhões de pares de bases ficou entre 0,06 e 1,67/mb, sendo que essa variação depende da definição dos atributos de um *hotspot*. Essa definição varia bastante entre os diversos autores.

Lefebvre & Labuda (2008) partem da premissa elaborada por Stephens (1986), de que a estimação da taxa de recombinação, a partir de dados genéticos de populações deve ser precedida de uma estimação do grau de informação relacionado ao segmento genômico a ser analisado. Conforme os autores "ao se considerar um conjunto de haplótipos de uma amostra de uma população, na qual todos os possíveis pares de haplótipos podem realizar recombinações recíprocas, de forma independente e aleatória, os possíveis resultados seriam: *i*) um haplótipo que não possui diferenças em relação aos haplótipos parentais; *ii*) um haplótipo filho, novo e diferente dos haplótipos parentais, porém, estruturalmente idêntico a um haplótipo já existente na amostra em análise; e *iii*) um haplótipo filho, novo e distinto dos haplótipos parentais e de qualquer outro haplótipo presente na amostra em estudo. Em outras palavras, a recombinação pode ser não informativa (produzindo haplótipos idênticos aos dos parentais) ou pode ser informativa". Decorre daí a definição das quantidades: fração de recombinações potencialmente informativas (*FIR*); fração de recombinações novas (*FNR*) e fração de recombinações recorrentes (*FBR - Fraction of Back Recombinations*). Esta última representa *crossover* recorrentes que são não informativos, quando se considera uma amostra de haplótipos de uma população.

Num conjunto de haplótipos que participam aleatoriamente de um processo de recombinação meiótica, assumindo-se uma mesma probabilidade de ocorrência de recombinação entre todos os pares de locos ao longo de uma sequência, *FIR* pode ser calculada pela seguinte expressão:

$$FIR = \sum_{i,j=1}^{k} \frac{f_i f_j D_{\max,ij}}{L},$$

Em que:

f_i e f_j : denotam as freqüências do *i*-ésimo e do *j*-ésimo haplótipo;

k : é o número de haplótipos, de modo que $\sum_{i,j=1}^{k} f_i f_j = 1$;

$D_{max,ij}$: é a distância entre os dois locos heterozigotos, maximamente afastados entre si, de um genótipo composto pelos haplótipos *i* e *j*;

L : é o tamanho do haplótipo.

Lefebvre & Labuda (2008) informam que, enquanto o *software* InfRec está computando a quantidade *FIR*, é mantida uma contagem de todas as recombinações que geram haplótipos idênticos a qualquer um já existente na amostra em análise. Essa contagem é usada para calcular *FBR* que, subtraída de *FIR*, fornece o valor da *FNR*. Enfim, *FNR* representa a proporção de recombinações que geram novos haplótipos na população amostrada.

$$FNR = FIR - FBR$$

Lefebvre & Labuda (2008) recorrem a Stephens (1986), para deduzir e apresentar a esperança matemática para os eventos de recombinação indetectáveis, *E(I)*.

$$E(I) = \frac{2\left(1 - e^{-\Theta}\right)}{\left(\Theta - e^{-\Theta}\right)}$$

Em que: $\Theta = 4N\mu$, sendo N o tamanho efetivo populacional e μ a taxa de mutação por segmento de DNA, por geração. Assim, *E(I)* de Stephen (1986) pode ser usada para calcular a esperança de *FIR*, *E(FIR),* pois *E(I)* = 1 – *E(FIR)*. O parâmetro Θ, segundo Tajima (1989), citado por Lefebvre & Labuda (2008), pode ser estimado pelo número médio de diferenças entre pares de locos numa amostra de haplótipos.

A probabilidade de ocorrência de genótipos que diferem entre si em *k* posições, em que *k = 0*, 1*, 2*, ..., obedece uma distribuição de Poison, pode ser obtida por:

$$P(k) = \frac{\left(e^{-\Theta}\Theta^{k}\right)}{k!}$$

As diferenças serão informativas somente se ocorrerem em duas ou mais posições, ou seja, $k \geq 2$. Entre essas posições podem ocorrer recombinações informativas na proporção (*k-1*)/(*k+1*). Daí, a esperança matemática para *FIR*, *E(FIR)*, ser igual a:

$$E\left(FIR\right) = \sum_{k=2}^{S} e^{-\Theta} \frac{\Theta^{k}\left(k-1\right)}{k!\left(k+1\right)}$$

Em que

S : é o número total de locos;

k : é o número de locos nos quais dois genótipos diferem entre si.

Para os casos em que os *croosover* são resolvidos pela via da conversão gênica, a expressão a seguir pode ser utilizada para calcular a esperança matemática da fração informativa da conversão gênica, *E(FIG)*:

$$E\left(FIG\right)=\sum_{k=2}^{S}e^{-\Theta}\frac{\Theta^{k}\left(k-1\right)}{k!\left(k\right)}\left(1-e^{-(k/L)t}\right)$$

Em que:

$1-e^{-(k/L)t}$: freqüência esperada de ocorrência de eventos de conversão gênica;

t : é o tamanho médio do fragmento em que ocorreu a conversão gênica.

A próxima etapa do método, após a estimação do grau informativo dos segmentos genômicos, é realizar a inferência dos haplótipos e do número mínimo de recombinações (R_{min}). Para a inferência dos haplótipos, seus propositores recomendam o uso do software PHASE, que implementa o método de Stephens et al. (2001). Para calcular o número mínimo de recombinações, a partir dos haplótipos, o *software* InfRec utiliza o método de Myers & Griffth (2003), implementado no *software* RecMin. Então, a taxa de recombinação populacional é estimada pela expressão:

$$\rho=\left(\frac{R_{\min}}{FNR\sum_{i=1}^{n-1}\left(1/i\right)}\right)$$

Lefebvre & Labuda (2008) advogam que esse método é rápido, simples e sem limitações para o tamanho da amostra e para a quantidade de locos. Sua relativa rapidez é devido ao fato de não realizar simulações, nem usar modelos baseados em *MCMC*, para recriar toda a história genealógica da sequência. A simplicidade é decorrente do fato de que a teoria subjacente é intuitiva, os cálculos são diretos, os resultados são transparentes e podem ser facilmente relacionados aos dados genéticos. Além disso, as esperanças matemáticas de *FIR* e *FIG* podem ser calculadas para ajudar analisar conjuntos particulares de dados, para os quais se interessa saber sobre as proporções relativas entre recombinação homóloga e conversão gênica.

3 MATERIAL E MÉTODOS

3.1 DADOS UTILIZADOS

Para a presente pesquisa foram utilizados dados da sequência de nucleotídeos do cromossomo 4 de *Arabidopsis thaliana*, disponíveis publicamente pelo projeto TAIR – The Arabidopsis Initiative Resource (CARNEGIE INSTITUTION FOR SCIENCE, 2009). Estes dados foram utilizados para realizarem-se buscas por elementos genômicos provavelmente associados à ocorrência de *crossover*.

Com o intuito de procurar por eventuais correlações entre a ocorrência de *crossover* e alguma propriedade estrutural do DNA foram utilizados dados das estimativas de intensidade de clivagem por radical OH^-. Estas estimativas foram calculadas para a posição de cada nucleotídeo da sequência do cromossomo 4 de *A. thaliana*., conforme metodologia proposta por Greenbaum et al. (2007).

Também foram utilizados dados de genotipagem de 41761 marcadores SNP, espalhados ao longo do cromossomo 4 de *A. thaliana* e avaliados em 362 acessos desta espécie (Nordborg, 2009). Estes acessos consistem em linhagens, cuja identificação e descrição foi elaborada por Li & Borevitz (2009). Trata-se de uma amostra que consiste num sub conjunto (*core*) não estruturado de acessos selvagens de *A. thaliana*. Esse tipo de dados foi utilizado para caracterizar a variação do desequilíbrio de ligação e da taxa de recombinação populacional ao longo do cromossomo 4 de *A. thaliana*.

A seleção dos 362 acessos de *A. thaliana* seguiu os seguintes procedimentos. Primeiramente, 6.389 acessos foram genotipados, utilizando-se 149 marcadores SNP. Quinze marcas tiveram falhas de leitura em mais de 20% dos acessos e 1.033 acessos tiveram falhas de genotipagem em mais de 20% das marcas. Todas as marcas e acessos que apresentaram problemas foram descartados, sobrando, então, 4.557 acessos e 134 marcas SNP. Em seguida, foram eliminados todos os acessos que eram idênticos entre si, provavelmente por coleta de descendentes de uma mesma planta. A análise de diversidade genética mostrou existirem pelo menos 1.749 haplótipos dentro das 4.557 linhagens remanescentes. Então, escolheu-se uma linhagem para representar cada haplótipo. Mesmo

após a ampliação de alguns haplótipos, ainda foram obtidos 1.589 grupos (Borevitz et al., 2009a). Diante desse quadro, foram selecionadas 384 linhagens que representavam 384 haplótipos maximamente divergentes entre si. Finalmente, considerando-se a disponibilidade de sementes e sugestões de outros colaboradores internacionais, elaborou-se a lista contendo 362 linhagens para a fase de genotipagem utilizando-se 250 mil marcas SNP (Borevitz et al., 2009b).

3.2 ESTIMATIVAS DE DESEQUILÍBRIO DE LIGAÇÃO

Para calcular as estimativas de desequilíbrio de ligação foram utilizados os dados genotípicos de 362 acessos de *A. thaliana*, utilizando-se 41761 marcadores SNP distribuídos ao longo do seu cromossomo 4. As estimativas de desequilíbrio de ligação para cada um dos 41760 pares consecutivos de locos foram calculadas utilizando-se a expressão elaborada por Hill & Robertson (1968):

$$\hat{r}^2_{AB} = \frac{(\hat{p}_{AB} - \hat{p}_A \hat{p}_B)^2}{\hat{p}_A \hat{p}_B \hat{p}_a \hat{p}_b}$$

Em que:

$\hat{r}^2_{AB}$: é a estimativa do coeficiente de determinação entre os locos *A* e *B*;

$\hat{p}_{AB}$: é a estimativa da frequência do haplótipo *AB* na população;

$\hat{p}_a$ ou $\hat{p}_b$: é a estimativa da frequência do alelo *a* ou *b* na população.

Cumpre observar que o termo $\hat{p}_{AB} - \hat{p}_A \hat{p}_B$ da expressão anterior equivale à estimativa da medida *D*, de desequilíbrio de ligação entre dois locos *A* e *B*, proposta por Lewontin & Kojima (1960).

Em seguida, foram realizados cálculos para se obterem estimativas de uma medida de alcance do desequilíbrio de ligação nos intervalos sucessivos entre os locos, utilizando-se a expressão:

$$\widehat{ALD}_{AB} = \frac{\hat{r}^2_{AB} \cdot d_{AB}}{0,2}$$

Em que:

$\widehat{ALD}_{AB}$: é a estimativa do alcance do desequilíbrio de ligação no intervalo entre os dois locos *A* e *B*;

$\hat{r}^2_{AB}$: é a estimativa do coeficiente de determinação entre os dois locos *A* e *B*;

d_{AB} : é a distância em pares de base (bp) entre o loco *A* e o loco *B*;

0,2 : valor crítico de referência para o cálculo do alcance do desequilíbrio de ligação.

3.3 ESTIMATIVAS DA TAXA DE RECOMBINAÇÃO POPULACIONAL

Para calcular as estimativas da taxa de recombinação populacional adotou-se o modelo proposto por Lefebvre & Labuda (2008). A escolha deste modelo em detrimento de outros, deu-se, principalmente, pelo motivo de custo computacional, sem perda da qualidade e da confiabilidade das estimativas calculadas. A abordagem apresentada por esses autores fornece estimativas da taxa de recombinação populacional com precisão aceitável, quando comparada a estimativas obtidas por outras modelagens, em conjuntos de dados comuns. Outro motivo que contribuiu para a decisão de adotar essa abordagem é o fato de o *software* elaborado por esses autores, para implementar o método, não ter limitações quanto ao tamanho da amostra (número de indivíduos genotipados), nem quanto ao número de marcas utilizadas. Quanto ao desempenho da metodologia para a detecção de *hotspots*, foram reportados resultados muito similares aos obtidos por outros métodos de detecção de *hotspots* de recombinação.

O modelo foi implementado em programa computacional denominado InfRec (Lefebvre & Labuda, 2008). O programa é escrito em linguagem Java e tem versões pré-compiladas para vários sistemas operacionais.

O arquivo de entrada para o programa InfRec tem um formato especial, igual ao formato dos arquivos "*.OUT", que consistem em saídas do programa PHASE (Stephens et al., 2001). Como os dados genéticos utilizados nas pesquisas do presente trabalho estavam gravados num arquivo da planilha Excel no formato "*.CSV", foi necessário fazer uma conversão desse formato para o formato "*.OUT". Esta conversão foi realizada através de programa construído especificamente para esse fim, em linguagem Java. O formato "*.OUT" tem a seguinte configuração:

As primeiras linhas do arquivos são as sequências contendo os alelos genotipados em cada posição. O número no início da linha identifica o indivíduo. No caso particular do arquivo utilizado neste trabalho, a primeira posição dessas linhas varia de um a 362.

1 TT...A 1.000000

2 TA...C 1.000000

... ...

362 GA...T 1.000000

O número do indivíduo é seguido pela sequência de letras que representa os alelos desse indivíduo em cada loco SNP genotipado. Neste caso, esta sequência contém 41761 letras. Em seguida, vem outro número com seis casas decimais, representando a frequência desse haplótipo na amostra a ser analisada. No presente trabalho todos os haplótipos foram, *a priori*, considerados como diferentes, com frequência igual a 1.000000 (correspondente a cada um dos 362 indivíduos). O programa InfRec avalia todos os haplótipos e, caso ocorra um haplótipo igual a outro, as frequências são automaticamente recalculadas. Os espaços presentes entre cada elemento nas linhas são imprescindíveis.

Após as linhas com os genótipos dos indivíduos segue-se:

Number of Individuals: 362

Positions of loci: 1470 1513 1575 … 18575378 18578708

"Number of individuals" consiste no número de indivíduos genotipados, neste caso 362; e "Positions of loci" é uma sequência de números representando as posições de cada loco, em pares de base. Neste caso, essa sequência conterá 41761 valores. Os números que indicam as posições devem ser separados por um espaço em branco.

Com o arquivo "*.OUT" preparado, procedeu-se à execução do programa InfRec para estimar a taxa de recombinação populacional (ρ). Convém ressaltar que, apesar do significado que a extensão do arquivo "*.OUT" possa ter para o leitor, essa extensão faz parte dos arquivos de entrada para o *software* InfRec.

O *software* InfRec realiza o cálculo das estimativas do número de eventos de recombinação que ocorrem em um dado fragmento genômico, ao longo da história da população amostrada, representada pelo estimador $\hat{\rho}$. Para calcular as estimativas $\hat{\rho}$ ao longo de uma região, o *software* utiliza janelas deslizantes para agrupar de dois a 100 locos de um fragmento genômico. Daqui por diante, as janelas deslizantes serão denominadas apenas de "janelas". O tamanho de uma janela (J) é determinado pela quantidade de locos que abrange, independentemente da densidade de locos no fragmento genômico abrangido. Como o tamanho das janelas é dado em número de locos e as distâncias entre estes variam com a densidade de marcas ao longo do fragmento genômico, ocorre que cada estimativa é associada a fragmentos de tamanhos distintos. Portanto, as estimativas $\hat{\rho}$ são convertidas e expressas em unidades de pares de base. Para isto, o *software* divide o valor da estimativa

$\hat{\rho}$ pelo comprimento (*L*), calculado da seguinte maneira:

$$L = \left(POS_{ultimo} - POS_{primeiro} \right) \left(\frac{J}{J-1} \right)$$

Em que:

L	: comprimento da janela em pares de bases;
POS_{ultimo}	: é a posição do último loco da janela;
$POS_{primeiro}$	: é a posição do primeiro loco da janela;
J	: é o tamanho da janela, em número de locos abrangidos.

Em seguida, calcula-se, então, a medida *ρ/Kb*:

$$\rho / Kb = \frac{\hat{\rho} / L}{10^3}$$

Doravante, a estimativa do número de eventos de recombinação, já corrigida para o tamanho do fragmento (*ρ/Kb*), será denominada simplesmente de $\boldsymbol{\rho}$.

A escolha do tamanho das janelas é relevante. No escopo do presente trabalho não foram feitas simulações para escolha de um tamanho ideal de janela. Apenas adotou-se a recomendação dos autores do método, utilizando-se janela de tamanho maior ou igual a oito locos. Essa recomendação advém de simulações realizadas pelos propositores do método, a partir das quais foi possível verificar que a variância das estimativas de $\boldsymbol{\rho}$ sofrem queda acentuada para tamanhos entre dois e seis locos, estabilizando-se a partir de sete a oito. Por outro lado, o uso de janelas muito amplas reduz a resolução com a qual se pode observar a variação da taxa de recombinação ao longo do segmento genômico. Assim, evitou-se utilizar janelas muito amplas. Enfim, foram realizadas estimativas de $\boldsymbol{\rho}$ para janelas com oito, dez, doze, dezesseis e vinte locos.

Para cada tamanho de janela foi gerado um arquivo de saída, contendo as estimativas de $\boldsymbol{\rho}$, entre outras. Foram gerados, então, cinco arquivos de saída, cada um contendo (41761 - *J* + 1) estimativas de $\boldsymbol{\rho}$. Por exemplo, para o arquivo referente à janela de tamanho igual a oito locos, existem 41761 - 8 + 1 = 41754 linhas com estimativas de $\boldsymbol{\rho}$. Para efeito de ilustração, um trecho de um dos arquivos de saída, contendo as dez primeiras linhas é apresentado na Tabela 1.

Dos arquivos de saída, foram utilizadas apenas as colunas "*Median*", "*Size*" e "*Rho/Kb*", que representam, respectivamente: o ponto médio entre as posições de um loco e o loco subsequente; o tamanho do segmento ao qual será ancorada a medida de $\boldsymbol{\rho}$; e a taxa de recombinação populacional já corrigida para o tamanho do segmento, $\boldsymbol{\rho}$.

Tabela 1. Trecho do arquivo de saída, gerado pelo *software* InfRec, contendo as estimativas de ρ, calculadas usando-se uma janela de tamanho igual oito locos.

WinSize	Median	Start	End	Size(bp)	RecMin	FIR	FBR	FNR	Rho/Kb
8	2225	1470	2980	1725	21	0,537	0,3124	0,2247	9,386
8	2258	1513	3003	1702	20	0,5524	0,1915	0,3609	5,6401
8	2610	1575	3645	2365	17	0,5119	0,1744	0,3375	3,6896
8	2724	1699	3749	2342	15	0,4396	0,1312	0,3084	3,5973
8	3149,5	1928	4371	2792	18	0,47	0,1785	0,2915	3,8317
8	6876,5	2062	11691	11004	13	0,404	0,061	0,343	0,5966
8	9120	2876	15364	14272	15	0,396	0,1959	0,2001	0,9097
8	10086	2980	17192	16242	16	0,4968	0,2938	0,203	0,8407
8	10342	3003	17681	16774	18	0,5135	0,2349	0,2786	0,6672
8	10897,5	3645	18150	16577	21	0,5127	0,2394	0,2733	0,8029

WinSize: tamanho da janela; *Median*: posição média, na qual se ancora as estimativas de ρ; *Start*, *End* e *Size*: posição de início, de término e tamanho do intervalo entre dois locos; *RecMin*: estimativa do número mínimo de *crossover* ocorridos no fragmento entre os dois locos; *FIR*: fração de recombinações informativas; *FBR*: fração de recombinações recorrentes; *FNR*: fração de novas recombinações; *Rho/Kb*: estimativa de ρ.

3.4 IDENTIFICAÇÃO DAS REGIÕES DE OCORRÊNCIA DE *HOTSPOTS* DE RECOMBINAÇÃO E DE LONGO ALCANCE DO DESEQUILÍBRIO DE LIGAÇÃO

O critério adotado para classificar um segmento de DNA como *hotspot* de recombinação foi baseado na seguinte metodologia. Primeiramente, foram elaboradas tabelas de freqüências das estimativas de ρ para cada um dos arquivos de saída; ou seja, para cada uma das janelas utilizadas. Na elaboração das tabelas foram utilizadas 1000 classes, de igual intervalo. Em seguida, calculou-se as frequências relativas acumuladas. O valor de ρ arbitrado como ponto de corte foi aquele correspondente a 99% das frequências relativas acumuladas. Então, todo intervalo cuja estimativa ρ era maior ou igual ao valor do ponto de corte foi considerado um *hotspot* de recombinação.

Após a classificação, os intervalos *hotspots* de cada arquivo foram destacados com uma cor diferente da dos demais intervalos, e, as cinco colunas de dados (uma para cada tamanho de janela) foram alinhadas entre si com base nas coordenadas das posições de referência no centro de cada intervalo. Fez-se, então, uma inspeção visual desde o 1º até o 41760º intervalo. Durante essa inspeção um determinado intervalo continuava sendo classificado como *hotspot* se, e somente se, esse intervalo tivesse sido classificado como *hotspot* em pelo menos quatro das cinco janelas (ou escalas) utilizadas. Mesmo nos casos em que ocorria uma sobreposição apenas parcial do intervalo de uma janela com o correspondente de outra, manteve-se a classificação de *hotspot* para o intervalo considerado.

Dessa forma, os intervalos considerados como *hotspots* de recombinação

formaram regiões contíguas. As coordenadas do início e do fim dessas regiões foram anotadas, a região recebeu um número, sendo a ela associada uma medida de ρ e outra de alcance do desequilíbrio de ligação. A medida de ρ associada a cada região *hotspot* foi calculada pela média aritmética das estimativas dos intervalos que compuseram a região, sempre tomadas na escala da maior janela utilizada. O mesmo procedimento foi adotado para calcular a média da medida de alcance do desequilíbrio de ligação. O resultado final consistiu numa tabela de seis colunas: número da região; posição inicial; posição final; tamanho da região em kilo pares de bases (kb); ρ médio dentro da região; e média dos valores do alcance do desequilíbrio de ligação dentro da região.

A mesma metodologia utilizada para identificar as regiões consideradas como *hotspots* de recombinação foi aplicada sobre as estimativas do alcance do desequilíbrio de ligação. Utilizaram-se os valores tão ou mais extremos que o valor correspondente a 99% das frequências relativas acumuladas como critério para classificar um intervalo como sendo de longo alcance. Fez-se a classificação em cada janela, alinharam-se as cinco colunas para a inspeção visual, mantendo-se a classe de intervalo com longo alcance apenas para aqueles intervalos cuja classificação se mantinha ao longo das diferentes escalas (ou janelas). O resultado final consistiu também de uma tabela de seis colunas contendo: número da região; posição inicial; posição final; tamanho da região em kb; alcance médio dentro da região; e ρ médio dentro da região.

Os dados das duas tabelas foram agrupados numa outra contendo as seis colunas, para todas as regiões identificadas e classificadas como *hotspots* ou como sendo de longo alcance, supostamente, assumidas como regiões *coldspots* de recombinação.

3.5 ANÁLISE DA DISTRIBUIÇÃO DE ELEMENTOS GENÔMICOS E ESTIMATIVAS DE PARÂMETROS ESTRUTURAIS

As coordenadas das posições iniciais das regiões identificadas como *hotspots* e das regiões com longo alcance do desequilíbrio de ligação foram tomadas como referência para extrair as sequências de cada região, do arquivo de sequências do cromossomo 4. Para extrair esse intervalo de nucleotídeos, a partir da sequência completa do cromossomo 4, foi elaborado um programa em linguagem Java. Para cada região *hotspot*, esse programa fez a leitura da sequência de DNA do cromossomo até chegar ao nucleotídeo cuja cooordenada era igual à posição inicial da região; daí, à medida que o programa fazia a leitura no

arquivo completo do cromossomo, simultaneamente era feita a gravação no arquivo de sequência da região, até chegar no nucleotídeo cuja coordenada era igual a da posição final da região. Assim, para cada região foi gerado um arquivo contendo a sequência de nucleotídeos do intervalo entre a posição inicial e a posição final da região considerada.

Usando-se este mesmo algoritmo, fez-se a leitura do arquivo contendo as estimativas de intensidade de clivagem por radical OH^- para o cromossomo 4, separando-se os valores dessas estimativas para cada região (tanto da classe *hotspot*, quanto da classe longo alcance). Assim, para cada região foi gerado um arquivo contendo a sequência de estimativas, por nucleotídeo, do intervalo entre a posição inicial e a posição final da região considerada. Para cada arquivo foi calculada a média aritmética de todas as estimativas de clivagem por radical OH^-. Essa média foi atribuída à região correspondente. Uma nova coluna de dados, contendo a média de clivagem por OH^-, foi, então, acrescentada às tabelas geradas pelo procedimento do item 3.4 do presente trabalho.

Após a individualização dos arquivos das sequências de nucleotídeos correspondentes a cada região da classe *hotspot* e da classe longo alcance, procedeu-se à contagem de alguns elementos genômicos dentro de cada arquivo: todas as possíveis sequências de di, tri e tetra nucleotídeos; o motivo CoHR identificado por Blumental-Perry et al. (2000); as sequências $(CCGNN)_{12}$ e $(CCGNN)_{48}$ identificadas por Kirkpatrick et al. (1999); o heptâmero 5'-ATGACGT-3', identificado por Steiner & Smith (2005a); e a sequência com 18 pares de bases 5'-GNVTATGACGTCATNBNC-3', identificada por Styeiner & Smith (2005b).

As informações referentes a cada elemento genômico geraram novas colunas para a tabela elaborada conforme procedimentos descritos no item 3.4 do presente trabalho. Para as contagens de elementos dentro de cada arquivo de sequências de nucleotídeos, foram realizadas as correções para o tamanho do arquivo, dividindo-se a frequência absoluta observada pelo número de pares de bases do fragmento, em kb.

3.6 ANÁLISE DAS CORRELAÇÕES DENTRO DOS FRAGMENTOS DAS CLASSES *HOT SPOT* E DE LONGO ALCANCE

Foram estimados coeficientes de correlação de Pearson entre todas as colunas da referida tabela, com dados de elementos genômicos, e as colunas contendo as médias de $\boldsymbol{\rho}$ e de $\widehat{ALD}$ para os fragmentos classificados ou como *hotspots* de recombinação ou como

regiões com longo alcance de desequilíbrio de ligação:

$$r = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(y)\,\mathrm{var}(x)}}$$

Os níveis de significância foram testados utilizando-se a estatística *t*, calculada pela expressão:

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}.$$

Como realizaram-se múltiplos testes de significância, aplicou-se o procedimento FDR (*False Discovery Ratio*), conforme metodologia de Benjamini & Hochberg (1995), sobre o conjunto ordenado de todos os *p*-valores, associados aos valores de *t*, referentes aos coeficientes de correlação estimados. O procedimento FDR foi aplicado duas vezes, de forma separada, uma para as correlações entre os elementos genômicos e a variável $\boldsymbol{\rho}$; outra vez para as correlações dos elementos genômicos com a variável $\widehat{ALD}$, alcance do desequilíbrio de ligação.

3.7 ANÁLISE DAS CORRELAÇÕES AO LONGO DO CROMOSSOMO 4 DE *A. thaliana*

3.7.1 Usando janelas deslizantes e sobrepostas

Por meio de programas de computador elaborados exclusivamente para esse fim, procedeu-se à contagem de elementos MAR/SAR (*Matrix Attachment Regions* / *Scafold Attachment Regions*) dentro de cada um dos 41760 intervalos entre os locos SNP ao longo do cromossomo 4 de *A. thaliana*. A contagem obtida dentro de cada intervalo (MARcont) foi transformada em frequência de elementos MAR/SAR por kb (MARfreq), dividindo-se a quantidade contada pelo tamanho do intervalo, em kb. Da mesma forma, procedeu-se para a variável porcentagem da soma das bases G e C (G+C%), em cada intervalo.

Para a variável intensidade de clivagem por radical OH^- (OH), o procedimento computacional lê e soma todos os valores das estimativas de clivagem por OH^- dentro de um intervalo entre dois locos. Em seguida, essa soma é dividida pelo número total de estimativas dentro do intervalo, obtendo-se uma média da intensidade de clivagem por

radical OH^- dentro de cada intervalo.

Os dados das estimativas do desequilíbrio de ligação e do alcance do desequilíbrio de ligação já tinham sido gerados na escala de pares de locos, pelo procedimento descrito no item 3.2.

Em resumo, os dados gerados consistiram em médias calculadas na escala de par de locos para as seguintes variáveis: intensidade de clivagem por radical OH^- (OH); frequência de elementos MAR/SAR por kb (MARfreq); contagem de elementos MAR/SAR por intervalo (MARcont); porcentagem média da soma das bases G e C por intervalo (G+C%); desequilíbrio de ligação entre locos adjacentes (LD); alcance do desequilíbrio de ligação por intervalo entre dois locos adjacentes (ALD).

Os dados das estimativas de cada uma dessas variáveis por par de locos consecutivos (escala igual a 2) foram sistematizados para escalas maiores, com os seguintes tamanhos de janelas: 6, 8, 10, 12, 16, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000 e 20000 locos. Utilizou-se a técnica das janelas deslizantes e sobrepostas, com passo de um em um loco. Esse procedimento é, de fato, o mesmo procedimento designado por médias móveis, com *lag* igual a um.

Para um determinado tamanho de janela (J), posiciona-se o centro da janela no loco $J/2$ e, em seguida, toma-se as medidas dos J-1 intervalos entre os J locos da janela, calculando-se uma média. O valor dessa média é atribuído ou ancorado na posição média entre o primeiro e o último loco que participa da janela. Em seguida, move-se a posição central da janela para o próximo loco ($J/2 + 1$) e repretem-se os procedimentos de cálculo da média e de sua ancoragem em uma posição. O procedimento de mover a janela se repete até o ponto em que o loco de maior posição da janela coincide com o último loco da série de dados.

Assim, para o conjunto de dados que contém 41761 locos, com 41760 intervalos, a quantidade de médias calculadas para a janela de tamanho 100, por exemplo, é 41660, obtida subtraindo-se 100 de 41760, pois o posicionamento da primeira janela se dará no loco de ordem 50 e o da última se dará no loco de ordem 41710.

Os números de graus de liberdade para aplicação do teste de t, associado ao coeficiente de correlação de Pearson, para cada tamanho de janela, foram considerados como: 41752, 41750, 41748, 41746, 41742, 41738, 41708, 41658, 41558, 41258, 40758, 39758, 36758, 31758 e 21758, respectivamente.

3.7.2 Usando janelas adjacentes e não sobrepostas

A sistematização dos dados em diferentes escalas, usando a técnica das janelas adjacentes não sobrepostas, foi realizada para o mesmo conjunto de variáveis descritas no item 3.7.1. Portanto, os dados iniciais para gerar as médias em escalas maiores são exatamente os mesmos; ou seja, as médias foram calculadas na escala de par de locos para as seguintes variáveis: intensidade clivagem por radical OH^- (OH); frequência de elementos MAR/SAR por kb (MARfreq); contagem de elementos MAR/SAR por intervalo (MARcont); porcentagem da soma das bases G e C por intervalo (G+C%); desequilíbrio de ligação entre par de locos adjacentes (LD); e alcance do desequilíbrio de ligação por intervalo entre dois locos adjacentes (ALD). A diferença desta etapa está no método de cálculo das médias em escalas subsequentemente maiores. Na técnica das janelas adjacentes e não sobrepostas, as medidas tomadas para o cálculo de uma média não participam do cálculo da média da posição adjacente, dentro de uma mesma escala.

Para uma escala com janelas de tamanho igual a duzentos locos, tomando-se o primeiro intervalo, que abrange a região do loco um ao loco de ordem duzentos, incluindo os intervalos de ordem um ao de ordem 199, calcula-se a primera média a partir das 199 primeiras estimativas da variável, em processo de sistematização. Essa média é atribuída ou ancorada na posição média entre a posição do primeiro loco e a do último loco participante da janela. Então, a janela se move duzentos locos adiante, de modo que a próxima média será calculada com as estimativas associadas aos intervalos de números 200 a 399. Daí as janelas serem denominadas por adjacentes e não sobrepostas. O procedimento é repetido até que o fim da série de dados seja alcançado.

Os tamanhos de janelas escolhidos, além do tamanho 2, em que estão os dados originais, são todos sub-múltiplos da quantidade total de locos, isto é: 3, 4, 5, 6, 8, 9, 10, 12, 15, 16, 18, 20, 30, 40, 48, 96, 144, 288, 522 e 1044 locos. Os graus de liberdade para o teste *t*, associado ao coeficiente de correlação de Pearson foram: 20878, 13918, 10438, 8350, 6958, 5218, 438, 4174, 3478, 2782, 2608, 2318, 2086, 1390, 1042, 868, 433, 288, 143, 78 e 38. Em seguida, procederam-se os cálculos dos coeficientes de corrrelação e dos respectivos testes *t* para cada par de variáveis em cada uma das escalas, protegendo-se como anteriormente descrito, o nível de significância do conjunto de testes pela aplicação do procedimento FDR, 5% e 1% de probabilidade.

4 RESULTADOS E DISCUSSÃO

4.1 CARACTERIZAÇÃO GERAL DO CROMOSSOMO 4 DE *Arabidopsis thaliana*

4.1.1 Dados descritivos da distribuição dos eventos de recombinação

A Tabela 2 apresenta algumas quantidades descritivas dos dados gerados acerca das estimativas de recombinação populacional (ρ), desequilíbrio de ligação (LD) e alcance do desequilíbrio de ligação (ALD).

Tabela 2. Algumas medidas de caracterização da recombinação populacional e do desequilíbrio de ligação ao longo do cromossomo 4 de *Arabidospsis thaliana.*

Variável[1]	Janela[2]	Mínimo	Máximo	Amplitude	Variância	Corte (99%)	Qtd.
ρ	8	0,000	166,251	166,251	89,941	38,794	768
	10	0,000	72,163	72,163	26,550	22,235	701
	12	0,149	41,804	41,654	13,569	16,331	685
	16	0,151	23,754	23,604	5,704	10,941	653
	20	0,164	16,399	16,235	3,542	8,727	640
LD	8	0,000	0,942	0,941	0,024	0,702	355
	10	0,004	0,940	0,936	0,021	0,666	363
	12	0,004	0,864	0,860	0,018	0,639	347
	16	0,005	0,814	0,809	0,015	0,600	311
	20	0,013	0,757	0,744	0,013	0,574	287
ALD	8	1,861	33.166,478	33.164,617	389.715,793	2.300,732	416
	10	2,701	26.251,508	26.248,807	320.545,114	2.126,422	440
	12	2,519	21.550,172	21.547,653	273.615,350	1.997,166	433
	16	7,213	16.119,139	16.111,926	214.474,989	1.817,199	433
	20	12,188	12.755,025	12.742,837	178.464,736	1.695,143	442

[1]: ρ representa a estimativa da taxa de recombinação populacional dividida por unidade de distância física ($\hat{\rho}$ /kb); LD: desequilíbrio de ligação; e ALD: alcance e desequilíbrio de ligação (kb); [2]: indica o número de marcas SNP utilizadas na estimação de ρ e de desequilíbrio de ligação; Mínimo e Máximo: representam, respectivamente, os valores mínimos e máximos para cada variável em cada janela utlizada; Corte (99%): consiste no valor arbitrado para se classificar regiões do cromossomo 4 como *hotspots* de recombinação ou alcances longos ; Qtd.: representa a quantidade de fragmentos cujas estimativas para ρ, LD e ALD apresentaram valores tão ou mais extremos que o valor Corte (99%).

Observa-se que a amplitude de variação das estimativas das medidas utilizadas diminui com o aumento da escala, ou seja, do tamanho da janela utilizada para computá-las. Isso está de acordo com o esperado, pois, com o aumento da escala as medidas tendem para a média do cromossomo e a variância tende para zero.

Os valores da taxa de recombinação populacional (ρ) e do alcance do

desequilíbrio de ligação (ALD) apresentam distribuições de frequências bastante assimétricas (Figuras 23 e 24, respectivamente), caracterizadas, de um lado, por um pico de concentração em torno das medidas próximas aos menores valores, e de outro, por um pequeno número de estimativas com valores extremos. Na Figura 23 pode ser observado que os valores de ρ, para a janela igual 20 SNP, se concentram entre 0,5 e 6, embora alguns poucos valores estejam próximos a 16.

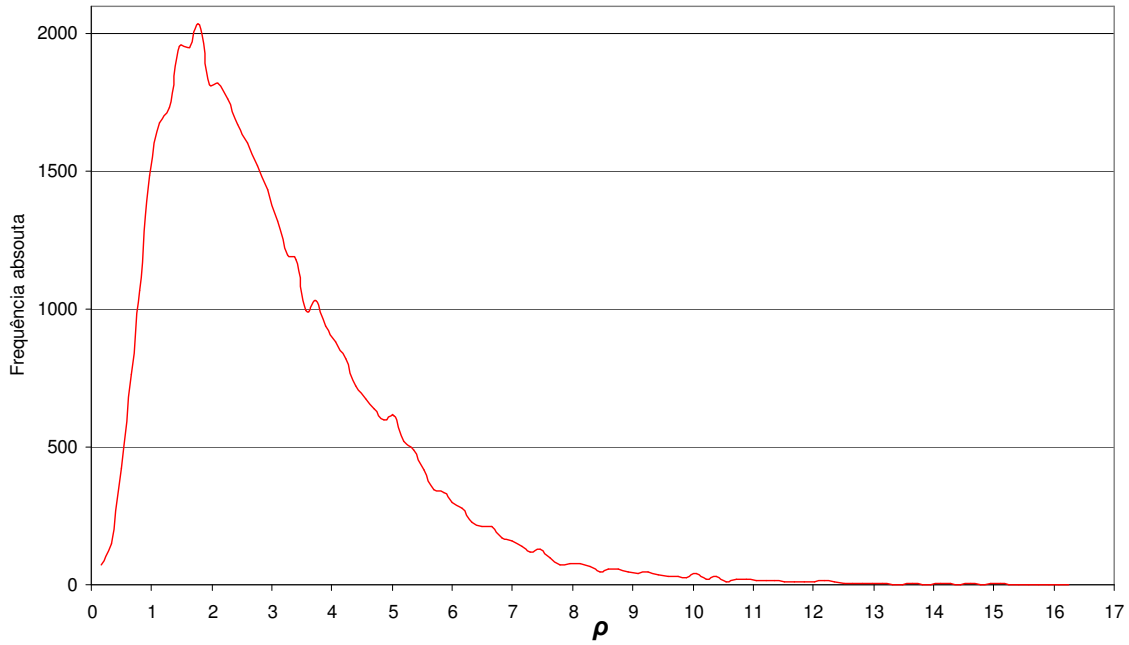


Figura 23. Distribuição de frequências das estimativas da taxa de recombinação populacional (ρ) ao longo do cromossomo 4 de *A. thaliana*, calculadas utilizando-se janelas de vinte SNP.

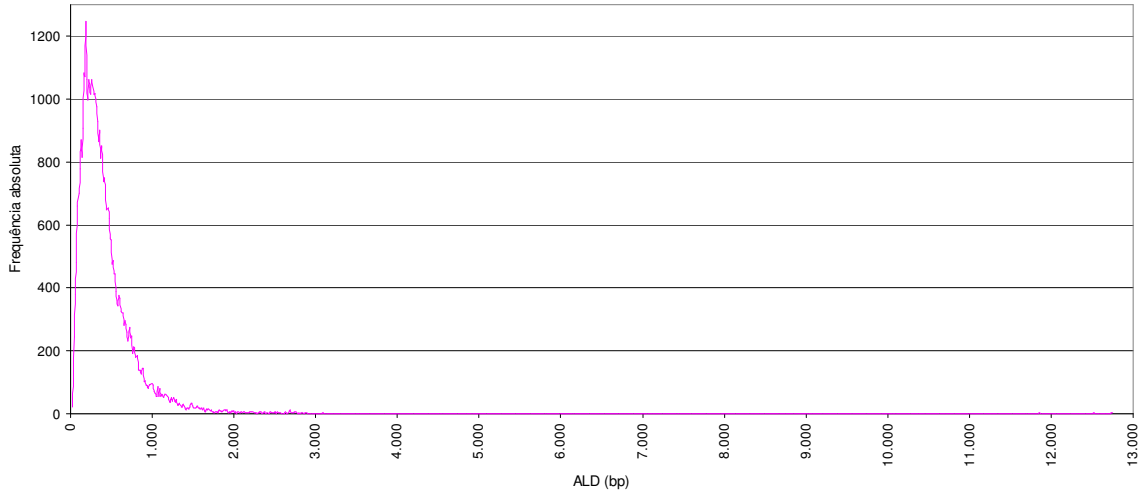


Figura 24. Distribuição de frequências das estimativas do alcance do desequilíbrio de ligação (ALD), em pares de bases (bp), ao longo do cromossomo 4 de *A. thaliana*, calculadas utilizando-se janelas de vinte SNP.

Conforme pode ser observado na Figura 24, a assimetria e a concentração dos valores das estimativas do ALD são ainda mais acentuadas. Enquanto a grande maioria dos valores se concentra entre cem e setecentos, observam-se algumas poucas estimativas com valores próximos a 13 mil.

Esses resultados são consistentes com aqueles apresentados por outros autores, para as medidas associadas à recombinação em *A. thaliana* (Drouaud et al., 2006) e no genoma humano (Myers et al., 2005). Convém ressaltar que esse padrão de distribuição das medidas de Rho e de ALD, que lembra as curvas de uma distribuição gama, é mantido em outras escalas, com o tamanho de janelas variando de 8 a 16. A Figura 16-A, da seção de revisão de literatura, extraída do trabalho de Myers et al. (2005), ilustra como a distribuição dos valores das medidas da taxa de recombinação no genoma humano, em várias escalas, também obedece um gráfico semelhante aos das Figuras 23 e 24.

Ao se analisar a relação entre a proporção do total de eventos de recombinação e a proporção do comprimento da sequência do cromossomo onde ocorreram os referidos eventos, verifica-se que a recombinação ocorre de forma bastante concentrada (Figura 25). Cinqüenta por cento dos eventos de recombinação ocorrem, aproximadamente, em apenas 13% do comprimento do cromossomo 4; aproximadamente 61% dos eventos de recombinação, quase dois terços do total, ocorrem em apenas 20% do comprirmento total do cromossomo; e ainda, 80% dos eventos de recombinação se concentram em pouco mais de um terço da sequência completa, aproximadamente 38%.

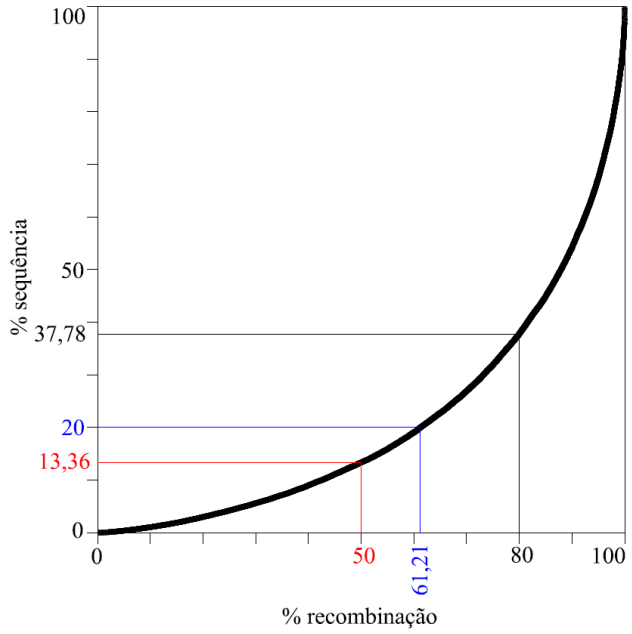


Figura 25. Proporção do total de eventos de recombinação em relação à proporção do comprimento do cromossomo 4 de *Arabidopsis thaliana*.

Esse padrão altamente concentrado de ocorrência dos eventos de recombinação também ocorre em todos os cromossomos do genoma de humanos, inclusive com uma concentração ainda mais acentuada (Myers et al., 2005). Apesar de pequena diferença entre os 23 cromossomos, na média, 80% da recombinação em seres humanos ocorre em 10% a 20% da sequência. Neste sentido, trabalhos realizados por Anderson et al. (2001, 2003), que quantificaram os eventos de recombinação na fase do zigóteno em células meióticas de tomateiro e de milho, mostraram o mesmo padrão de concentração; ou seja, a grande maioria dos eventos ocorriam em pequenas proporções do comprimento dos cromossomos. Também em milho, estudo de Tenallion et al. (2002) revelou o mesmo padrão de distribuição dos eventos de recombinação ao longo do cromossomo 1. Não há, portanto, casos relatados até o presente momento que indiquem um padrão de distribuição regular dos eventos de recombinação ao longo do comprimento dos cromossomos.

4.1.2 Identificação de *hotspots* de recombinação e de fragmentos de DNA com longo alcance do desequilíbrio de ligação

Após a aplicação dos critérios de identificação de fragmentos de DNA que poderiam ser considerados como *hotspots* de recombinação ou como regiões caracterizadas pela presença de longos alcances do desequilíbrio de ligação, foram identificados 85 fragmentos classificados na primeira categoria e 41 fragmentos, cuja média dos valores de alcance do desequilíbrio de ligação é maior ou igual ao do ponto de corte arbitrado (Tabela 2). Os dados relativos aos fragmentos da classe *hotspots*, suas posições e os valores médios de taxa de recombinação populacional (ρ) e alcance do desequilíbrio de ligação (ALD) estão sistematizados na Tabela 3.

Tabela 3. Fragmentos do cromossomo 4 de *Arabidopsis thaliana* classificados como *hotspots* de recombinação por terem apresentado valores de ρ tão ou mais extremos que o ponto de corte arbitrado.

Número do Fragmento	Posição (bp)		Tamanho (kb)	ρ médio	ALD (bp)
	Inicial	Final			
1	967.785	968.757	0,972	12,1	36
2	1.652.474	1.655.961	3,487	9,3	81
3	1.779.716	1.783.577	3,861	9,8	107
4	2.046.126	2.049.029	2,903	11,6	48
5	2.055.120	2.056.034	0,915	11,3	100
6	2.171.463	2.173.448	1,985	8,9	89
7	2.249.217	2.250.719	1,502	9,9	115
8	2.253.995	2.255.689	1,694	9,5	133
9	2.259.532	2.261.109	1,578	12,7	113
10	2.271.911	2.273.196	1,285	10,1	68

continua ...

Tabela 3. Continuação

Número do Fragmento	Posição (bp) Inicial	Final	Tamanho (kb)	ρ médio	ALD (bp)
11	2.572.914	2.573.526	0,612	8,8	65
12	2.579.599	2.580.826	1,227	9,6	91
13	2.613.558	2.614.389	0,831	9,1	82
14	2.765.699	2.766.482	0,783	11,2	225
15	2.770.932	2.771.975	1,043	11,6	104
16	2.785.479	2.786.850	1,371	9,1	125
17	3.142.147	3.143.270	1,123	9,1	177
18	3.755.753	3.756.734	0,981	11,6	115
19	4.612.028	4.612.550	0,522	11,6	175
20	4.628.354	4.631.426	3,072	12,6	198
21	4.718.381	4.720.210	1,829	8,8	191
22	4.729.961	4.732.880	2,920	9,5	96
23	4.744.314	4.746.042	1,729	9,4	100
24	4.763.388	4.764.503	1,115	12,0	111
25	5.205.989	5.209.141	3,152	11,6	57
26	5.264.576	5.267.546	2,970	9,9	183
27	5.375.032	5.376.377	1,345	12,3	316
28	5.624.744	5.626.910	2,167	9,2	62
29	5.659.134	5.660.456	1,323	9,6	65
30	5.968.581	5.970.210	1,630	9,3	104
31	6.019.125	6.019.776	0,652	8,9	359
32	6.130.585	6.132.659	2,074	9,7	46
33	6.193.867	6.195.769	1,903	11,3	56
34	6.232.321	6.233.127	0,806	8,8	255
35	6.718.382	6.719.564	1,183	11,6	92
36	6.827.208	6.828.963	1,755	9,7	143
37	6.914.427	6.917.185	2,758	11,6	110
38	6.963.996	6.968.447	4,452	12,0	138
39	6.980.201	6.982.183	1,982	9,3	64
40	6.983.843	6.986.247	2,404	11,0	65
41	7.005.239	7.007.876	2,637	10,2	90
42	7.011.099	7.011.642	0,543	12,2	92
43	7.148.420	7.151.373	2,954	11,2	75
44	7.196.331	7.197.307	0,977	9,0	76
45	7.200.942	7.203.262	2,320	10,9	132
46	7.207.793	7.210.210	2,417	9,4	70
47	7.222.772	7.223.327	0,555	11,4	36
48	7.227.022	7.229.458	2,436	10,3	216
49	7.273.153	7.277.982	4,830	11,4	68
50	7.317.471	7.319.669	2,198	13,4	14
51	7.333.004	7.337.513	4,509	12,8	138
52	7.607.118	7.608.861	1,743	11,1	136
53	7.858.280	7.859.394	1,114	12,3	127
54	7.946.903	7.948.553	1,650	9,0	118
55	7.969.474	7.971.251	1,777	9,2	248
56	8.001.226	8.003.670	2,444	10,9	79
57	8.007.565	8.012.439	4,875	9,9	43
58	8.025.420	8.026.683	1,264	9,7	128
59	8.280.044	8.281.436	1,393	9,5	173
60	8.296.735	8.298.664	1,929	10,6	72
61	8.317.699	8.319.355	1,656	12,2	58
62	8.539.442	8.539.808	0,366	11,3	236
63	9.042.497	9.044.420	1,923	9,9	83
64	9.093.648	9.096.485	2,837	9,4	176
65	9.411.726	9.414.000	2,274	11,5	109

continua ...

Tabela 3. Continuação

Número do Fragmento	Posição (bp) Inicial	Final	Tamanho (kb)	ρ médio	ALD (bp)
66	9.427.131	9.430.856	3,725	10,6	210
67	9.455.212	9.457.283	2,071	10,0	163
68	9.466.465	9.469.387	2,922	10,8	59
69	9.471.370	9.475.090	3,720	10,6	88
70	9.479.461	9.481.780	2,320	9,9	53
71	9.526.062	9.529.852	3,790	11,9	117
72	9.545.690	9.547.674	1,985	9,9	75
73	9.560.354	9.562.091	1,737	9,8	94
74	9.581.535	9.582.195	0,660	11,1	31
75	9.591.813	9.594.070	2,257	9,4	253
76	9.691.321	9.692.545	1,224	14,0	259
77	9.740.515	9.741.967	1,453	9,5	191
78	9.800.187	9.802.051	1,864	9,2	150
79	9.924.110	9.925.835	1,725	11,1	169
80	10.230.710	10.231.889	1,179	11,6	43
81	10.450.668	10.453.156	2,488	9,6	207
82	10.906.113	10.907.151	1,038	9,1	118
83	12.728.043	12.729.809	1,766	9,3	377
84	13.077.012	13.079.243	2,231	13,9	133
85	17.094.533	17.095.015	0,482	11,6	60
-	-	Média	1,955	10,5	124
-	-	Mínimo	0,366	8,8	14
-	-	Máximo	4,875	14,0	377
-	-	Amplitude	4,509	5,2	363

Os dados relativos aos fragmentos com valores extremos para o alcance do desequilíbrio de ligação, suas posições e os valores médios de taxa de recombinação populacional (ρ) e alcance do desequilíbrio de ligação (ALD) estão sistematizados e apresentados na Tabela 4.

Tabela 4. Fragmentos do cromossomo 4 de *Arabidopsis thaliana* classificados como regiões de alcances longos por terem apresentado valores de alcance do desequilíbrio de ligação (ALD) tão ou mais extremos que o do ponto de corte arbitrado.

Número da Região	Posição (bp) Inicial	Final	Tamanho (kb)	ρ médio	ALD (bp)
1	7.625	16.902	9,3	0,542	2.103
2	23.107	31.176	8,1	0,890	2.196
3	963.192	966.606	3,4	2,441	2.140
4	1.683.697	1.721.918	38,2	0,450	2.697
5	2.929.608	2.957.135	27,5	0,303	12.405
6	3.013.035	3.036.104	23,1	0,583	1.955
7	3.080.556	3.106.111	25,6	0,714	2.541
8	3.218.327	3.282.538	64,2	0,252	2.436
9	3.982.144	4.048.429	66,3	0,251	2.504
10	4.100.692	4.112.014	11,3	1,269	1.886
11	4.241.184	4.246.357	5,2	1,199	2.259
12	4.272.893	4.285.480	12,6	1,274	2.123
13	4.465.053	4.493.833	28,8	1,067	3.146
14	4.515.758	4.531.506	15,7	1,419	2.063
15	4.566.194	4.586.436	20,2	0,835	1.987

continua ...

Tabela 4. Continuação

Número da Região	Posição (bp) Inicial	Final	Tamanho (kb)	ρ médio	ALD (bp)
16	4.675.848	4.689.965	14,1	0,786	1.717
17	5.786.155	5.791.634	5,5	1,537	1.812
18	5.869.999	5.905.081	35,1	0,445	2.585
19	7.418.322	7.423.886	5,6	1,797	1.972
20	11.372.108	11.386.045	13,9	1,295	1.801
21	11.777.443	11.783.464	6,0	1,281	1.920
22	12.787.330	12.792.750	5,4	0,940	1.892
23	12.897.627	12.907.053	9,4	1,015	2.102
24	13.616.610	13.631.643	15,0	0,885	2.046
25	13.876.168	13.893.197	17,0	1,309	2.274
26	14.356.769	14.382.602	25,8	0,751	2.020
27	14.425.953	14.434.580	8,6	1,284	1.942
28	14.709.238	14.713.432	4,2	0,598	2.168
29	14.945.364	14.953.193	7,8	0,877	1.851
30	15.211.461	15.220.510	9,0	0,804	1.921
31	15.518.741	15.552.486	33,7	0,526	2.219
32	15.808.309	15.823.231	14,9	0,874	1.848
33	16.029.189	16.042.302	13,1	0,706	1.972
34	16.481.725	16.489.604	7,9	1,211	1.813
35	16.553.353	16.559.556	6,2	1,110	1.732
36	16.774.282	16.795.125	20,8	0,626	2.122
37	17.357.863	17.371.087	13,2	1,818	1.909
38	17.823.615	17.845.478	21,9	0,742	2.298
39	17.911.028	17.927.657	16,6	0,665	2.336
40	18.193.508	18.200.342	6,8	1,290	1.856
41	18.412.690	18.419.156	6,5	1,243	1.777
-	-	Média	17,2	0,973	2.350
-	-	Mínimo	3,4	0,251	1.717
-	-	Máximo	66,3	2,441	12.405
-	-	Amplitude	62,9	2,190	10.688

Os fragmentos da classe *hotspots* ocorreram com o dobro da frequência dos fragmentos da classe de alcances longos. Existe uma região de aproximadamente 4 mb no cromossomo 4 de *A. thaliana*, entre as posições 7,4 mb e 11,4 mb, na qual não houve a incidência de qualquer fragmento da classe de alcances longos. Metade dos fragmentos dessa classe estão na parte proximal do cromossomo, entre 0 mb e 7,4 mb e a outra metade está na parte distal, entre 11,4 mb e 18,5 mb. Coincidentemente, 32 dos 85 fragmentos da classe *hotspots* recaíram, justamente, sobre a região entre 7,4 mb e 11,4 mb, caracterizando-a como uma longa região com alcances relativamente baixos do desequilíbrio de ligação. Assim, uma caracterização em escala mais ampla poderia separar o cromossomo 4 de *A. thaliana* em três partes: os terços distal e proximal, menos recombinogênicos, contendo vários fragmentos caracterizados por possuírem valores relativamente altos para o alcance do desequilíbrio de ligação; e o terço central mais recombinogênico, contendo relativamente mais fragmentos da classe *hotspots* de recombinação.

Um aspecto que distingue claramente os fragmentos da classe *hotspots* dos da classe de alcances longos é a diferença entre seus tamanhos. Os fragmentos da classe de alcances longos têm um tamanho médio de 17,2 kb, variando de 3,4 kb a 66,3 kb. Por outro lado, os fragmentos da classe *hotspots* têm tamanho médio de 1,9 kb, variando de 0,4 kb a 4,9 kb. Por alguma razão, as altas frequências de ocorrências de eventos de recombinação não se estendem por longas distâncias no cromossomo, elas tendem a ocorrer em fragmentos comparativamente curtos de sequências do DNA. Já os fragmentos que se caracterizam como de alcances longos são, em média, dez a 60 mil vezes maiores que os fragmentos da classe *hotspots*, caracterizando-se também por manter o comportamento deste alcance por distâncias relativamente grandes ao longo do cromossomo.

Na comparação da posição dos fragmentos de cada classe e das intensidades de suas respectivas medidas, com a posição de intervalos estudados por Drouaud et al. (2006), também no cromossomo 4 de *A. thaliana*, verificam-se coincidências confirmadoras. Um exemplo é o fragmento de número 2, da classe dos *hotspots* nas análises do presente trabalho, o qual coincide com o *hotspot* de número 7 do trabalho de Drouaud et al. (2006). Por outro lado, a região entre os intervalos 59 a 63, no estudo de Drouaud et al. (2006), apresenta-se como um vale de baixíssimas taxas de recombinação, enquanto nesta pesquisa ocorreu a incidência dos fragmentos de números 3 a 18 na classe de *hotspots*. Como Drouaud et al. (2006) genotiparam apenas 71 SNP ao longo do cromossomo 4, diferentemente dos 41761 utilizados nas análises do presente estudo, fica claro que a natureza mais heterogênea da ocorrência de recombinação nessa região escapou ao baixo poder de detecção da análise daqueles autores. Apesar disso, outras coincidências, como a do intervalo 42 de Drouaud et al. (2006) e o fragmento 55 do presente estudo, mostram que o padrão geral da curva de variação da recombinação é confirmado pelo padrão presente nos resultados apresentados até o momento.

4.1.3 Variação da taxa de recombinação, do desequilíbrio de ligação e do alcance do desequilíbrio de ligação

A região entre 1,8 mb e 4,2 mb, por conter o centrômero e uma região heterocromática, denominada *knob*, mereceu uma análise mais detalhada. O gráfico da Figura 26 mostra a variação das estimativas do alcance do desequilíbrio de ligação (ALD) ao longo dessa região. O pico na medida de ALD, centrado na posição próxima a 2,95 mb, foi resultante da existência de uma estimativa de desequilíbrio de ligação alto (0,856) entre

as marcas das posições 2.917.573 e 2.967.968, afastadas entre si em 50.395 pares de bases, localizadas nas proximidades da parte distal da região centromérica. Esta estimativa de ALD com valor extremo, de quase 216.000 bp, influencia as estimativas de vários fragmentos vizinhos em todas as escalas utilizadas para as análises do presente trabalho. Quanto maior a janela utilizada para os cálculos, maior é a quantidade de fragmentos afetados. Para a janela igual a 8 SNP ocorre influência das posições 2.939.564 a 2.949.711. Para a janela igual a 20 SNP, por exemplo, a influência se dá da posição 2.935.525 bp até a 2.950.932 bp. Considerou-se essa medida como um "*outlier*", pelo fato de que a sua magnitude prejudica a apresentação visual das demais, em todas as escalas utilizadas.

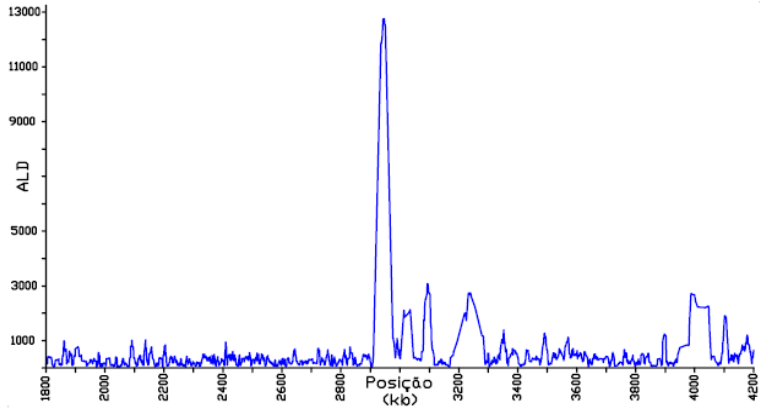


Figura 26. Variação do alcance do desequilíbrio de ligação (ALD) ao longo da região entre 1,8 mb e 4,2 mb, no cromossomo 4 de *Arabidopsis thaliana*. Estimativas de ALD calculadas com janela de 20 SNP.

Nas listas das estimativas para cada tamanho de janela fez-se, então, o descarte das medidas atribuídas aos fragmentos afetados por este "*outlier*". Assim, a configuração do padrão da variação de ALD nessa região passou a ser representada como ilustrado na Figura 27.

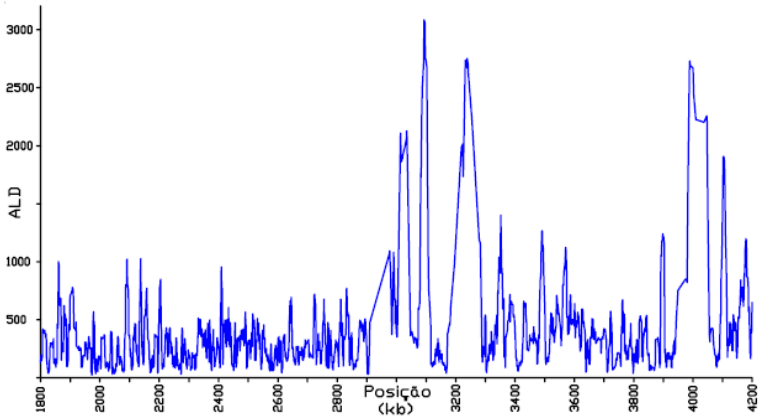


Figura 27. Variação do alcance do desequilíbrio de ligação (ALD) ao longo da região entre 1,8 mb e 4,2 mb, no cromossomo 4 de *A. thaliana*. Estimativas de ALD calculadas com janela de 20 SNP, excluindo-se o "outlier", estimativa de ALD entre as posições 2.917.573 bp e 2.967.968 bp.

Daqui por diante, fica explicitado que todas as demais figuras, contendo alguma informação sobre o alcance do desequilíbrio de ligação, foram elaboradas sem a presença da medida "*outlier*" anteriormente mencionada.

A variação da taxa de recombinação populacional (***ρ***) ao longo do cromossomo 4 de *A. thaliana*, com tamanho de janela igual a 8 SNPs, é apresentada na parte superior da Figura 28. É possível detectar regiões que consistem em vales de supressão da recombinação, como as das posições próximas de 2,8 mb, 3,2 mb , 3,9 mb, 5,8 mb, 11,1 mb e 15,6 mb, nas quais ocorrem picos nos valores de alcance do desequilíbrio de ligação (ALD). De forma geral o terço distal do cromossomo 4, a partir da posição 11 mb, apresenta valores e amplitudes menores na variação de ***ρ*** e, consequentemente, valores maiores e razoável variação na amplitude de ALD.

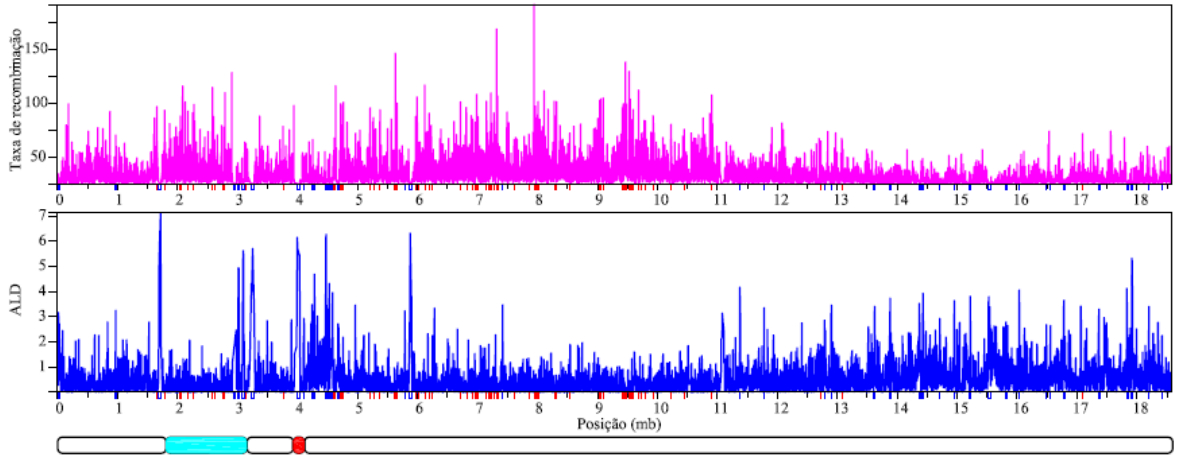


Figura 28. Variação da taxa de recombinação populacional (***ρ***) e do alcance do desequilíbrio de ligação (ALD) ao longo do cromossomo 4 de *Arabidopsis thaliana*, calculadas usando uma janela igual a 8 SNP. Na parte superior, em magenta, a variação de ***ρ*** (número de *crossover* por kb). Na parte inferior, em azul, a variação de ALD (kb). Junto à parte inferior do eixo x encontram-se pequenos retângulos que representam os fragmentos com valores tão ou mais extremos que os valores de Corte (99%) arbitrados: os azuis representam fragmentos com valores extremos de ALD; e os vermelhos representam os fragmentos com valores extremos de ***ρ***. Na base da figura encontra-se uma representação esquemática do cromossomo 4 de *A. thaliana*, na mesma escala do eixo x. A parte vermelha representa a região centromérica e, a ciano o *knob* heterocromático.

Ao se observar a parte inferior da Figura 28, que apresenta a variação de ALD ao longo do cromossomo 4, verifica-se que as regiões do centrômero e uma região heterocromática, denominada *knob*, são ladeadas por picos no alcance do desequilíbrio de ligação. Porém, a região interna do *knob* heterocromático mostra uma alternância entre picos e vales locais de recombinação.

Conforme identificado e caracterizado por Fransz et al. (2000), o *knob* heterocromático é composto, predominantemente, por elementos de transposição; mas inclui alguns genes. A região de aproximadamente 1,5 mb, incluindo o *knob*, é invertida quando se compara o acesso Columbia (Col) com o Landsberg (Ler) de *A. thaliana* e, esse fato, em conjunto com os efeitos da presença do próprio *knob*, podem consistir numa provável explicação para a ocorrência de considerável supressão da recombinação nos flancos dessa região.

Os gráficos da Figura 28 parecem indicar uma coincidência entre as posições de valores extremos de ALD com as de supressão da recombinação. Porém, quando analisados em mais detalhes, os gráficos mostraram que em várias regiões as linhas azuis e magenta sobem ou descem juntas. O coeficiente de correlação de Pearson para este par de variáveis é de r = -0,39 (p-valor < 10^{-9}), mas o gráfico da Figura 29 é que apresenta uma relação mais clara entre ALD e $\boldsymbol{\rho}$. Valores extremos de uma das variáveis nunca estão associados com valores extremos da outra. Esta constatação valida, de certa forma, a estratégia de se ter utilizado os valores extremos de ALD para identificar *coldspots* de recombinação.

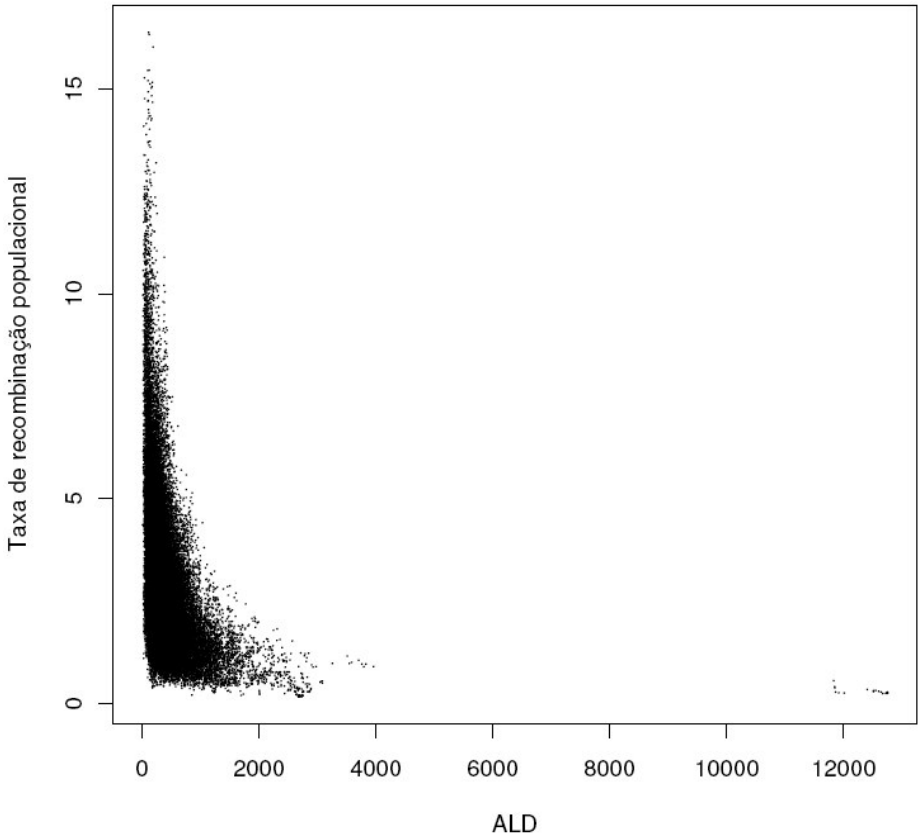


Figura 29. Gráfico de dispersão entre estimativas de alcance do desequilíbrio de ligação (ALD) a de taxas de recombinação populacional ($\boldsymbol{\rho}$) ao longo do cromossomo 4 de *Arabidopsis thaliana*, calculadas usando uma janela igual a 20 SNP.

Outra região que foi analisada com mais detalhes fica entre as posições 1 bp e 182466 bp. A atenção particuar a essa região se deve ao fato de que esta corresponde ao intervalo número 70, no trabalho de Drouaud et al. (2006). Nessa pesquisa, este mostrou-se como o *hotspot* mais intenso de todos os intervalos analisados e, por isso, eles o dividiram em 15 partes para mapear, com maior resolução, os eventos de recombinação. Os autores detectaram uma distribuição não homogênea dos eventos de recombinação dentro do intervalo 70, identificando picos localizados de concentração de *crossover*. Com base em leitura visual, os quatro *hotspots* de recombinação de Drouaud et al. (2006) se localizam, aproximadamente, nas posições 70 kb, 120 kb, 140 kb e 160 kb a partir da extremidade proximal. O *hotspot* da posição 120 kb é o menos intenso dos quatro, e o da posição 160 kb o mais intenso. Na resolução obtida no presente estudo, foi possível confirmar não só a presença dos quatro *hotspots* nas posições estimadas por Drouaud et al. (2006), como também as mesmas proporções nas diferenças de suas respectivas intensidades. Essa comparação está ilustrada na Figura 30, na qual as setas vermelhas marcam os quatro *hotspots* identificados e confirmados.

Mézard (2006) publicou uma figura mais detalhada, elaborada a partir dos dados da pesquisa de Drouaud et al. (2006), extendendo a caracterização da taxa de recombinação até a posição 800 kb. Além dos hotspots identificados por Drouaud et al. (2006), a caracterização realizada por Mézard (2006) permitiu identificar pelo seis *hotspots* e tres *coldspots* de recombinação, entre as posições 200 kb e 800 kb. Os *coldspots*, destacados em ciano na Figura 30, se posicionam, respetivamente, nos intervalos de 200 kb a 230 kb, 310 kb a 360 kb e 570 kb a 630 kb. Os *hotspots* identificados pelo autor, destacados em magenta na Figura 30, estão centrados, respectivamente, nas posições 260 kb, 400 kb, 510 kb, 670 kb, 700 kb e 740 kb. Observa-se pequenas discrepâncias no posicionamento e nas intensidades relativas dos *coldspots* e *hotspots* caracterizados por Mézard (2006), em relação à caractrização resultante da presente pesquisa. A magnitude dessas discrepências podem ser atribuídas às diferenças nas resoluções utilizadas nas duas pesquisas. De forma geral, há uma concordância aceitável no posicionamento e nas intensidades relativas dos *hotspots*. O padrão geral da variação de $\boldsymbol{\rho}$, ao longo dos primeiros 800 kb do cromossomo, foi captado de forma satisfatória pelas pesquisas de Drouaud et al. (2006) e Mézard (2006). Os resultados do presente trabalho corroboram os daqueles autores, somente ampliando o nível de resolução com o qual se descreveu a variação de $\boldsymbol{\rho}$, neste intervalo.

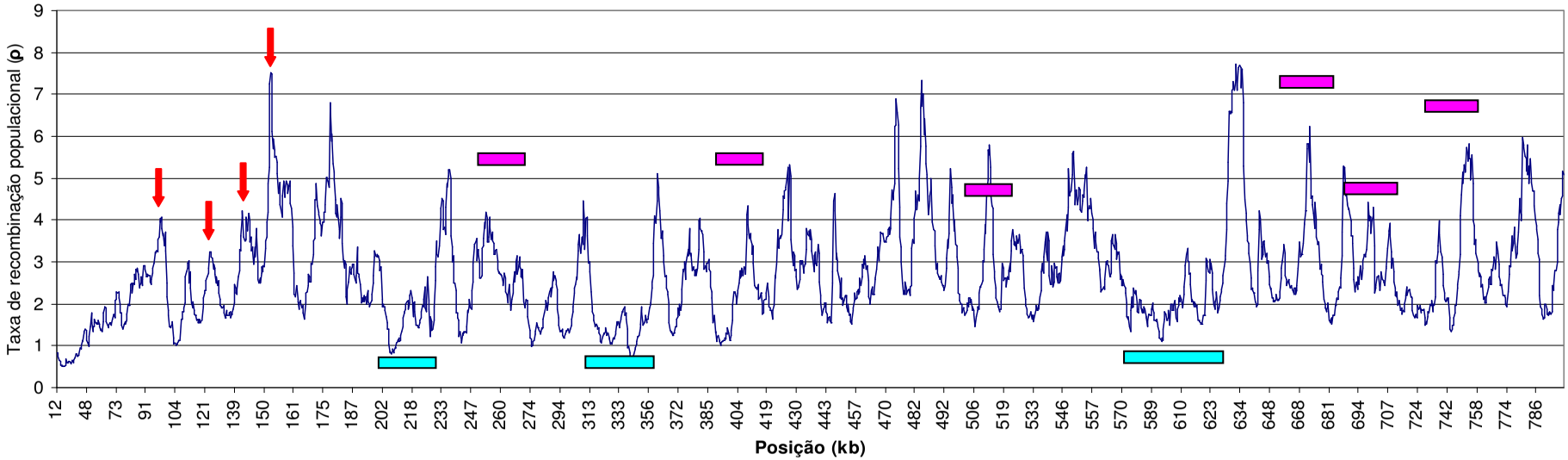


Figura 30. Variação da taxa de recombinação populacional (***ρ***) ao longo dos primeiros 800 kb do cromossomo 4 de *Arabidopsis thaliana*. As setas vermelhas indicam as posições de quatro *hotspots* identificados por Drouaud et al. (2006). As barras magenta e ciano representam, respectivamente, as regiões *coldspots* e *hotspots* de recombinação caracterizadas por Mézard (2006). As diferenças nas alturas das barras magenta representam as diferenças de intensidade dos *hotspots* caracterizados por Mézard (2006).

No contexto da comparação feita anteriormente, convém retomar as contribuições do trabalho e Allard (1963). A recombinação é um processo que está sobre controle genético e a métrica que a quantifica, a taxa de recombinação, pode ser estimada a partir de alguns tipos de populações segregantes. A taxa de recombinação varia enormemente pela influência de fatores ambientais, com o *background* genético, a idade das plantas, a época de florescimento, entre épocas de plantio e, dentro das épocas, varia entre as plantas. Drouaud et al. (2006), para realizarem sua pesquisa genotiparam 752 plantas F2 resultantes de um cruzamento entre os acessos Columbia e Landsberg de *A. thaliana*. Considerando as observações de Allard (1963), poder-se-ia inquirir sobre a aplicabilidade de estudos que se propõem a caracterizar os denominados *hotspots* de recombinação, atribuindo-lhes uma posição, intensidade e tamanho em pares de base. Alguns estudos de sintenia apontam para uma não conservação dos *hotspots*, afirmando que são fenômenos efêmeros e que, portanto, não mereceriam grande atenção das pesquisas.

Por outro lado, apesar de Drouaud et al. (2006) terem utilizado uma população diferente de *A. thaliana* e um nível de resolução bastante distinto do empregado no presente trabalho, esses autores conseguiram identificar um padrão para a variação da taxa de recombinação ao longo do cromossomo 4. Assim, caracterizaram vários *hotspots*, informando sua posição e fornencendo uma medida de suas respectivas intensidades. As posições e as intensidades desses *hotspots* de recombinação têm coincidência razoável com as obtidas pelas análises realizadas no presente trabalho, que fizeram uso de 41761 marcadores moleculares do tipo SNP, sobre uma amostra de 362 indivíduos da espécie. A ação do processo de recombinação parece deixar sua marca no genoma. Se estas marcas são efêmeras na escala de tempo evolutiva, podem não o ser na escala de tempo que interessa a outros tipos de estudos e aplicações.

O melhoramento genético de plantas, por exemplo, pode fazer uso dessa informação para aprimorar o processo de escolha e determinação da densidade de marcas moleculares ao longo dos cromossomos, para estudos de associação. Uma alternativa seria o estabelecimento de densidade de marcas de forma proporcional à intensidade da recombinação: quanto maior a taxa de recombinação, maior a densidade de marcas, e vice-versa. Ao invés de se buscar por uma distribuição eqüidistante das marcas, a referida alternativa proporcionaria uma distribuição homogênea quanto aos valores de desequilíbrio de ligação (LD) entre pares de marcas. Para um conjunto fixo de marcas a serem utilizadas,

a distribuição homogênea quanto ao LD proporcionaria maior grau de informatividade por marca em relação à que seria obtida por uma distribuição eqüidistante de marcas. Enfim, levar em consideração a heterogeneidade da variação de ρ ao longo dos cromossomos possibilita captar maior parte da variação genética do genoma do que a que seria captada ao se adotar uma distribuição eqüidistante entre marcas. Fica a ressalva de que é necessário elaborar estudos mais detalhados para verificar diferenças da caracterização de ρ entre populações.

4.1.4 Análise da distribuição de elementos genômicos em diferentes escalas

Além de uma variação detectável ao longo de um cromossomo, numa dada escala, os valores das estimativas da taxa de recombinação populacional (ρ), de desequilíbrio de ligação (LD) e de alcance do desequilíbrio de ligação (ALD) também sofrem influência da escala na qual se realizam os cálculos. Ao se analisar as informações descritivas apresentadas na Tabela 2, foi possível observar o efeito da escala sobre a magnitude e a variação das estimativas.

A Figura 31 apresenta os gráficos do padrão da variação de ρ , LD e ALD ao longo do cromossomo 4 de *A. thaliana*, utilizando-se janela com 20 SNP. Figuras semelhantes, para as escalas correspondentes aos tamanhos de janela iguais a 10 SNP, 12 SNP e 16 SNP, encontram-se nos Apêndices A-1, A-2 e A-3, respectivamente. A forma pela qual a variação da ocorrência de eventos de recombinação se comporta ao longo do cromossomo, observando-se a Figura 31, na escala de 20 SNP, permite observar que há vales intensos de recombinação nas posições aproximadas de 1,7 mb, 2,8 mb, 3,2 mb, 3,9 mb, 5,8 mb, 11,1 mb e 15,6 mb e que o terço distal, a partir da posição 11,5 mb, tem valores de ocorrências de eventos de recombinação flutuando em torno de uma média menor do que a observada para as demais partes do cromossomo. Projetando-se uma linha imaginária a partir dos pontos que representam os vales de recombinação até a parte inferior da Figura 31, é possível verificar a coincidência dos vales de recombinação com picos elevados dos valores de ALD. Os gráficos da Figura 31 e os das Figuras dos anexos I, II e III proporcionam praticamente a mesma informação visual, embora as janelas para as escalas de 10 SNP e 20 SNP correspondam, respectivamente, a fragmentos genômicos com tamanhos de 4,5 kb e 9,0 kb, em média.

A generalização promovida pelo uso de médias calculadas com base em agre-

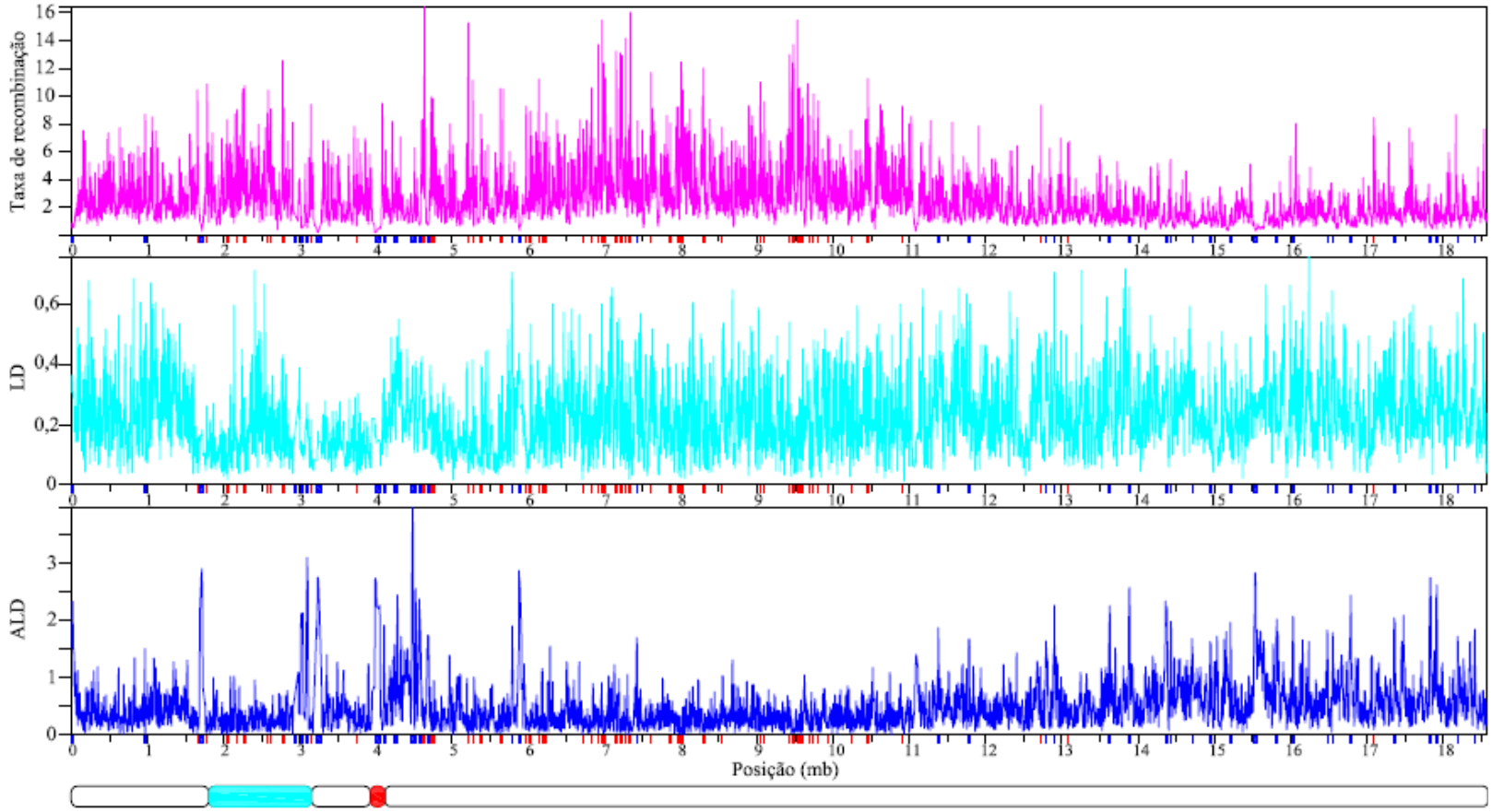


Figura 31. Variação da taxa de recombinação populacional (***ρ***), desequilíbrio de ligação (LD) e alcance do desequilíbrio de ligação (ALD) ao longo do cromossomo 4 de *Arabidopsis thaliana*. Na parte superior, em magenta, a variação de ***ρ***. Ao centro, em ciano, a variação de LD e, na parte inferior, em azul, a variação de (ALD). As estimativas de ***ρ*** , LD e ALD foram calculadas usando uma janela com 20 SNP. Junto à parte inferior do eixo x encontram-se pequenos retângulos que representam os fragmentos com valores tão ou mais extremos que os valores de Corte (99%) arbitrados: os azuis representam fragmentos com valores extremos de ALD; e os vermelhos representam os fragmentos com valores extremos de ***ρ***. Na base da figura encontra-se uma representação esquemática do cromossomo 4 de *A. thaliana*, na mesma escala do eixo x. A parte vermelha representa a região centromérica e, a ciano o *knob* heterocromático.

gação de valores de uma escala, para outra mais ampla, pode ser visualmente observada em escala de algumas dezenas de kilo pares de bases. A inspeção visual dos gráficos da Figura 32 permite observar que o efeito qualitativo do aumento da escala é uma tendência de estabelecimento de uma linha mais suavizada, que representa as variações "regionais" da variável em estudo. As ondulações dessa linha mais suavizada representam uma provável periodicidade de variação dos dados na escala observada. Nesse caso particular, observando-se a escala de 100 SNPs, o período parece ter tamanho de 1,5 mb. Tomando como ponto de partida a posição de 1,5 mb, é possível observar picos subsequentes dos valores do alcance do desequilíbrio de ligação nas posições 3,0 mb, 4,5 mb, 6,0 mb, 7,5 mb e assim por diante. Mas, a busca por padrões de variação com recursos da inspeção visual pode ser em muito melhorada por técnicas que têm sido incorporadas à análise de variação de elementos genômicos mais recentemente. Entre estas, cita-se a técnica de análise *wavelet*.

Li & Holste (2004) estudaram a variação de alguns elementos genômicos no cromossomo 21 de seres humanos. Os autores, utilizando a técnica de análise *wavelet*, verificaram que o conteúdo GC apresenta oscilações periódicas, com altos e baixos em uma periodicidade de 500 kb. A técnica *wavelet* parece consistir em um recurso de análise com boas perspectivas de aplicação na modelagem da variação da taxa recombinação e de elementos genômicos ao longo de sequências de DNA. Essa técnica enriquece a informação visual obtida a partir dos gráficos, pois proporciona uma caracterização quantificada das variações de um elemento genômico ao longo do cromossomo, em várias escalas.

Para as estimativas de alcance do desequilíbrio de ligação (ALD), o efeito da generalização em escalas mais amplas, com as janelas variando de 20 SNP a 10.000 SNP, pode ser visualizado na Figura 32. Uma analogia que poderia ser aplicada à análise do que ocorre nesses gráficos seria a do que ocorre em sistematização de dados censitários. O gráfico correspondente à escala de 20 SNPs representaria os dados tomados em cada setor censitário, em geral, algumas quadras de um bairro. O gráfico correspondente à escala de 50 SNPs representaria um distrito que agrega vários setores censitários. Assim por diante, as escalas subsequentemente superiores representariam um bairo, uma cidade, um município, um estado, uma região do país, um país, um continente, etc.

Feições de variação que são observáveis em escalas menores, como, por exemplo, os dois picos de variação do alcance do desequilíbrio de ligação observados nas

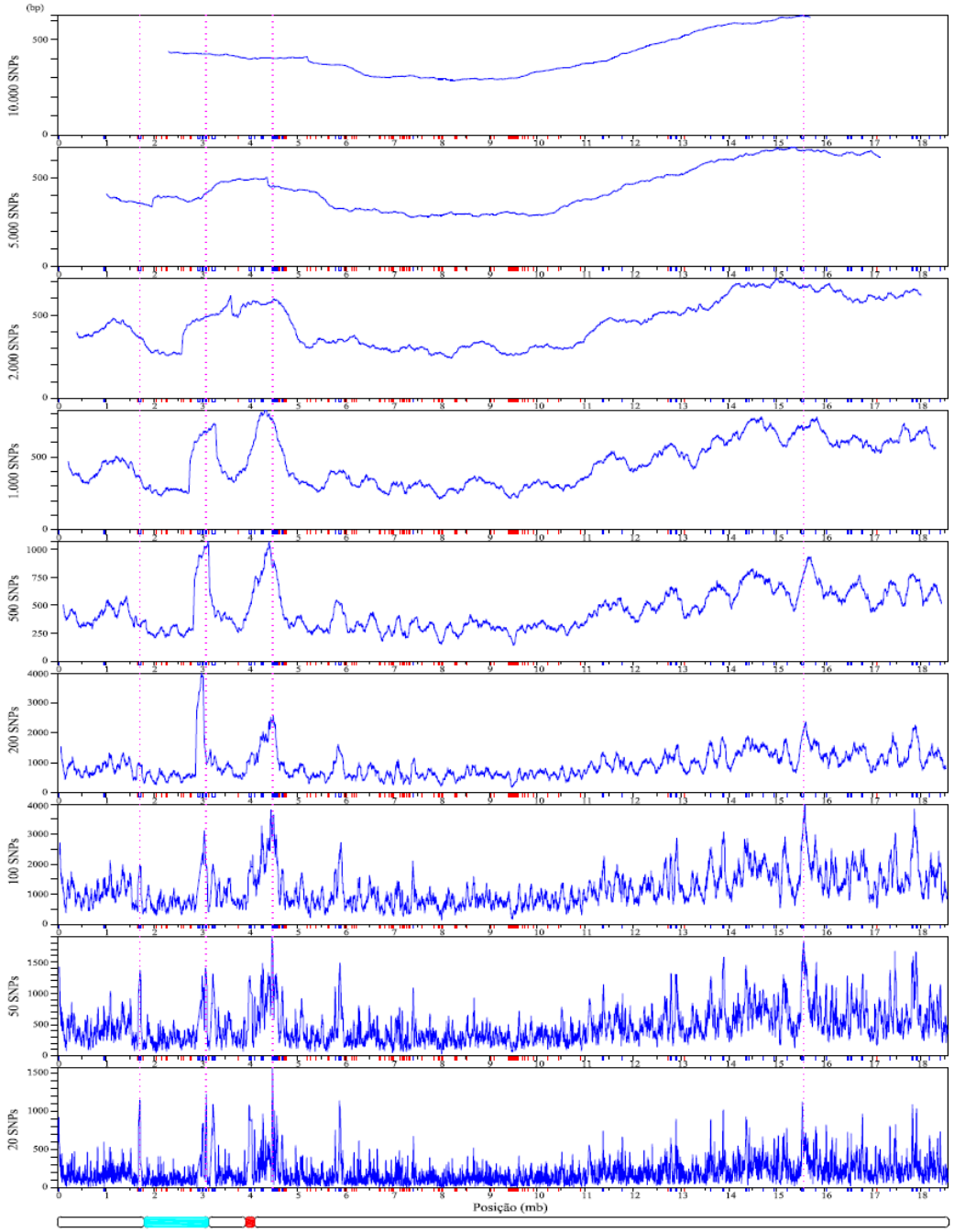


Figura 32. Variação do alcance do desequilíbrio de ligação (ALD) ao longo do cromossomo 4 de *Arabidopsis thaliana*. De baixo para cima, os gráficos correspondem, respectivamente, às escalas cujas janelas têm tamanho de 20, 50, 100, 200, 500, 1000, 2000, 5000 e 10000 SNPs. As linhas pontilhadas verticais servem para ancorar as mesmas posições do cromossomo nos gráficos de todas as escalas. Junto à parte inferior do eixo x encontram-se pequenos retângulos que representam os fragmentos com valores tão ou mais extremos que os valores de Corte (99%) arbitrados: os azuis representam fragmentos com valores extremos de ALD; e os vermelhos representam os fragmentos com valores extremos de $\boldsymbol{\rho}$. Na base da figura encontra-se uma representação esquemática do cromossomo 4 de *A. thaliana*, na mesma escala do eixo x. A parte vermelha representa a região centromérica e, a ciano o *knob* heterocromático.

posições de 3,0 mb e 4,5 mb, ao longo das escalas de 20 SNPs a 1.000 SNPs se convertem em apenas uma grande elevação que se extende de 2,6 mb a 5,2 mb, nas escalas maiores. Essa perda na representação da variação, quando se sistematizam dados de uma escala para outra, pode causar perda no poder de detecção de uma provável correlação entre variáveis em estudo. O contrário também poderia ser dito: correlações não detectáveis em determinada escala poderão ser detectadas em outras. Uma discussão mais detalhada sobre esses aspectos será feita mais adiante no tópico 4.2.

Neste ponto convém ressaltar que, no contexto de análise de dados genômicos, uma técnica de análise preliminar pode ser derivada da simples observação visual de gráficos que caracterizam a variação de determinado elemento genômico ao longo de um cromossomo, em várias escalas. Essa técnica consistiria em identificar a escala máxima na qual feições importantes da variação de determinado elemento ainda estariam mantidas. Por exemplo, ao se observar os gráficos da Figura 32 identificam-se os dois picos nos valores do alcance do desequilíbrio de ligação (entre 3,0 mb e 4,5 mb), separados por um vale de valores menores. Essa feição se mantém até a escala de 1.000 SNPs, mas é perdida na escala de 2.000 SNPs. Se, num exemplo hipotético, para determinado estudo fosse importante manter a informação da variação em ALD dentro da região de 3,0 mb a 4,5 mb, a recomendação seria utilizar escalas de no máximo 1.000 SNPs agregados por janela.

Em síntese, da análise dos gráficos multi-escalas, elaborados para as variáveis utilizadas no presente estudo constatou-se que o limite de escala ideal para as análises de determinada variável pode não o ser para outra, considerando-se o critério de se manter feições bem definidas em escalas menores. Para exemplificar, apresenta-se na Figura 33 os gráficos multi-escala para a variação das estimativas da intensidade de clivagem por radicais OH^- ao longo do cromossomo 4 de *A. thaliana*. Neste caso, a escala máxima seria a de cem SNPs ou duzentos SNPs. A perda de informação é bastante acentuada quando se sistematizam os dados da escala de duzentos para a de quinhentos SNPs.

Adimitindo-se, portanto, que a escala de sistematização dos dados influencia as possíveis associações entre variáveis genômicas, é de se esperar que, para cada par de variáveis, existirá uma escala na qual a correlação é máxima. Isto foi investigado para algumas variáveis, sendo tratado na próxima seção.

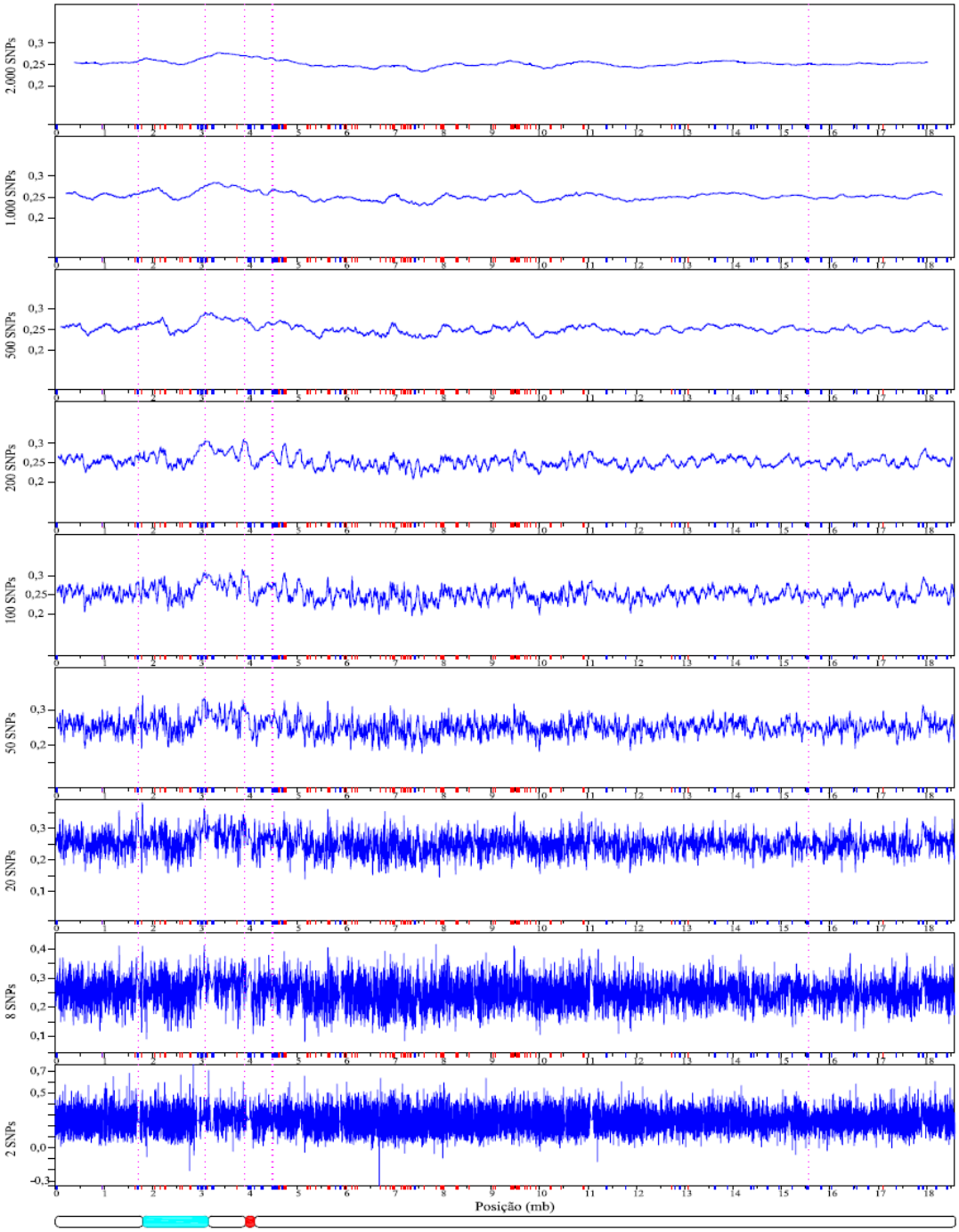


Figura 33. Variação da intensidade de clivagem por radicais $OH^{\cdot}$ ao longo do cromossomo 4 de *Arabidopsis thaliana*. De baixo para cima, os gráficos correspondem, respectivamente, às escalas cujas janelas têm tamanho de 20, 50, 100, 200, 500, 1000, 2000, 5000 e 10000 SNPs. As linhas pontilhadas verticais servem para ancorar as mesmas posições do cromossomo nos gráficos de todas as escalas. Junto à parte inferior do eixo x encontram-se pequenos retângulos que representam os fragmentos com valores tão ou mais extremos que os valores de Corte (99%) arbitrados: os azuis representam fragmentos com valores extremos de ALD; e os vermelhos representam os fragmentos com valores extremos de $\boldsymbol{\rho}$. Na base da figura encontra-se uma representação esquemática do cromossomo 4 de *A. thaliana*, na mesma escala do eixo x. A parte vermelha representa a região centromérica e, a ciano o *knob* heterocromático.

4.2 ANÁLISES DE CORRELAÇÃO

4.2.1 Análises de correlação entre a distribuição de elementos genômicos e as taxas de recombinação e de desequilíbrio de ligação dentro das classes de fragmentos

No escopo do presente trabalho buscou-se encontrar possíveis associações entre alguns elementos e/ou variáveis genômicas e as sequências dos fragmentos das duas classes arbitradas e utilizadas nas análises, *hotspots* de recombinação e fragmentos com valores extremos do alcance do desequilíbrio de ligação (*coldspots*). As variáveis e elementos genômicos que participaram dessas análises foram: frequências de certas combinações de nucleotídeos em sequências de dois a quatro pares de bases; porcentagem da soma das bases G e C por fragmento (G+C%); alguns motivos relacionados na literatura como sendo associados a *hotspots* de recombinação; contagem de elementos estruturais MAR/SAR (MARcont); frequência de MAR/SAR por kb (MARfreq); frequência do pentanucleotídeo CCGNN; média da intensidade de clivagem por radicais OH^- (ClivOH).

Das 336 combinações de bi, tri e tetra nucleotídeos, apenas 54 mostraram correlação significativa com a média das estimativas do alcance do desequilíbrio de ligação (ALD); assim como a variável MARcont, de acordo com o teste de *t*-Student, protegido pelo procedimento FDR para o nível de significância de 5%. Destas 54, apenas 21 apresentaram correlação significativa por FDR a 1% de probabilidade (Tabela 5). Todas as sequências que apresentaram correlação significativa com as médias do alcance do desequilíbrio de ligação são tri ou tetra nucleotídeos: AAAG, AACT, AAG, AAGA, AAGC, ACCG, AGAA, AGCT, AGTA, ATAC, ATCG, ATGC, CAA, CAAA, CCGG, CGTC, CTAA, CTAG, CTCT, GAAA, GAAC, GACT, GAGT, GATC, GCAA, GCGT, GCTA, GGAA, GGAT, GTA, GTAT, GTCA, GTGG, GTGT, GTTA, GTTG, TAAG, TACA, TAG, TAGA, TAGT, TCAA, TCCC , TCCG, TCT, TCTC, TCTG, TCTT, TGAG, TGC, TGCC, TGCT, TGGA e TTAG.

Na análise de correlações entre elementos genômicos e a taxa de recombinação populacional ($\boldsymbol{\rho}$), por fragmento, 18 elementos mostraram correlações estatisticamente significativas a 5% (Tabela 6), sendo que, destes, apenas quatro mantiveram suas correlações significativas a 1% de probabilidade. São eles: a contagem de elementos MAR/SAR por fragmento (MARcont), a frequência do motivo CCGNN e as frequências dos tetra nucleotídeos CAAA e TGGA.

Tabela 5. Correlações entre certos elementos genômicos e as médias das estimativas de alcance do desequilíbrio de ligação, dentro dos 126 fragmentos das classes *hotspot* e *coldspot*.

EG[1]	*r*	p-valor	FDR	EG[1]	*r*	p-valor	FDR
pGCTA	0,4808	$< 10^{-7}$	**	pCTAA	0,2085	0,0191	*
pAAGC	0,4771	$< 10^{-7}$	**	pCCGG	-0,2072	0,0199	*
pCAAA	0,4546	$< 10^{-7}$	**	pGAAA	-0,2071	0,0200	*
MARcont	0,4381	$< 10^{-6}$	**	pGCG	-0,2053	0,0211	NS
pTAAG	0,3870	$< 10^{-5}$	**	pGTGT	0,2050	0,0213	*
pAAAG	0,3868	$< 10^{-5}$	**	pCTCT	-0,2024	0,0230	*
pAGTA	0,3786	$< 10^{-4}$	**	pCGTC	-0,2016	0,0236	*
pGTAT	0,3769	$< 10^{-4}$	**	pTGCT	-0,2014	0,0238	*
pAGAA	0,3736	$< 10^{-4}$	**	pTC	-0,2003	0,0246	NS
pTAGA	0,3644	$< 10^{-4}$	**	pATGT	0,1994	0,0252	NS
pGTTA	0,3273	0,0002	**	pGATC	-0,1988	0,0257	*
pCTAG	0,3269	0,0002	**	pCCGC	0,1982	0,0261	NS
pGACT	0,3268	0,0002	**	pTCCG	-0,1959	0,0279	*
pAGCT	0,3219	0,0002	**	pCGCA	0,1948	0,0288	NS
pAAG	0,2969	0,0007	*	pGAAC	0,1944	0,0292	*
pGCAA	0,2952	0,0008	**	pAGCG	-0,1942	0,0293	NS
pTAG	0,2925	0,0009	**	pACGA	-0,1939	0,0296	NS
pGAGT	0,2809	0,0014	**	pCATT	-0,1937	0,0298	NS
pCAA	0,2772	0,0017	*	pATGG	-0,1933	0,0301	NS
pGTCA	0,2761	0,0017	**	pACTA	0,1933	0,0301	NS
pAAGA	0,2744	0,0019	**	pTAGT	0,1932	0,0302	*
pTCT	-0,2725	0,0020	**	pGCGT	-0,1928	0,0305	*
pATAC	0,2708	0,0022	**	pAACC	0,1901	0,0330	NS
pTGC	-0,2613	0,0031	*	pGG	-0,1897	0,0334	NS
pTTAG	0,2472	0,0053	**	pCGAA	-0,1887	0,0343	NS
pAACT	0,2384	0,0072	*	pATC	-0,1885	0,0345	NS
pGTA	0,2348	0,0081	*	pGGAT	-0,1874	0,0356	*
pCGT	-0,2331	0,0086	NS	pATCT	-0,1869	0,0361	NS
pATGC	-0,2301	0,0095	*	pAGA	0,1861	0,0369	NS
pACCG	0,2282	0,0102	*	pTCAA	0,1845	0,0386	*
pTGGA	-0,2272	0,0105	*	pTCTT	-0,1844	0,0387	*
pAA	0,2250	0,0113	NS	pCGA	-0,1844	0,0388	NS
pCTA	0,2243	0,0116	NS	pCGTG	-0,1840	0,0392	NS
pTCTC	-0,2207	0,0130	*	pGGA	-0,1817	0,0417	NS
pAGT	0,2187	0,0139	NS	pCGGG	-0,1806	0,0430	NS
pGTGG	-0,2158	0,0152	*	pTGG	-0,1801	0,0437	NS
pTCCC	0,2154	0,0154	*	pTCTG	-0,1798	0,0440	*
pTGCC	-0,2151	0,0156	*	pGCCT	-0,1795	0,0444	NS
pTACA	0,2147	0,0158	*	pCG	-0,1786	0,0454	NS
pGGAA	-0,2143	0,0160	*	pTCGT	-0,1781	0,0461	NS
pATCG	-0,2131	0,0166	*	**...**	**...**	**...**	**...**
pAACG	-0,2118	0,0173	NS	pGATA	0,0014	0,9876	NS
pTGAG	0,2116	0,0174	*	pGTAC	-0,0011	0,9898	NS
pGTTG	-0,2094	0,0186	*	pAGGC	0,0010	0,9909	NS

[1]: os elementos genômicos (EG) que participaram da análise de correlações foram os seguintes: todas as sequencias de bi, tri e tetra nucleotídeos (pAA, PCC, pGG, pTT, pAC, pAG, pAT, ..., pTTTT); a porcentagem da soma das G e C por fragmento (G+C%); a freqüência dos motivos CCGNN; a intensidade de clivagem por radical OH^- (ClivOH); a quantidade de elementos MAR/SAR por fragmento (MAR cont); a freqüência de elementos MAR/SAR por kb em cada fragmento (MAR freq). A tabela foi truncada entre as linhas de ordem 85 e 338, pois, da linha 82 em diante, todas as correlações foram não significativas.

r: coeficiente de correlação de Pearson.

FDR: * e ** correspondem a valores significativos a 5% e 1%, respectivamente; NS corresponde a valores não significativos.

Uma associação qualitativamente importante foi a que se apresentou entre as

contagens de elementos do tipo MAR/SAR (MARcont) dentro dos 126 fragmentos analisados. A contagem total de elementos MAR/SAR foi 115. Destes, houve apenas 18 elementos MAR/SAR que incidiram em 14 dos 85 fragmentos classificados como *hotspots* de recombinação. As outras 97 ocorrências de elementos MAR/SAR incidiram em 30 dos 41 fragmentos classificados como *coldspots*. E, a maioria desses 97 elementos MAR/SAR está concentrada nos fragmentos da parte proximal e distal do cromossomo. O coeficiente de correlação entre MARcont e ALD é positivo, r = 0,438 (Tabela 5), e negativo com $\boldsymbol{\rho}$, r = -0,548 (Tabela 6). O padrão da variação de ALD observado no gráfico da escala de 5.000 SNPs, da Figura 32, mostra a tendência de aumento de ALD do centro para as extremidades do cromossomo. Esse conjunto de constatações explicita a associação entre a variação de ALD e a contagem de elementos MAR/SAR ao longo do cromossomo, em escala mais ampla.

Tabela 6. Correlações entre certos elementos genômicos e as médias das estimativas da taxa de recombinação populacional ($\boldsymbol{\rho}$), dentro dos 126 fragmentos das classes *hotspot* e *coldspot*.

EG[1]	*r*	p-valor	FDR	EG[1]	*r*	p-valor	FDR
MARcont	-0,5475	$< 10^{-10}$	**	pTCCG	0,1955	0,0282	*
pTGGA	0,2752	0,0018	**	pGCCG	0,1954	0,0283	*
pCAAA	-0,2533	0,0042	**	pGAT	0,1913	0,0319	NS
pCCGG	0,2455	0,0056	*	pTCGT	0,1885	0,0346	*
pCGT	0,2424	0,0062	*	pTAAG	-0,1861	0,0369	*
pCCGNN	0,2381	0,0073	**	pGTGG	0,1848	0,0383	*
pTGG	0,2292	0,0098	*	pAATT	-0,1843	0,0388	NS
pCG	0,2256	0,0111	NS	pAGTA	-0,1832	0,0400	NS
pCCG	0,2243	0,0116	NS	pCGGG	0,1808	0,0428	NS
ClivOH	0,2183	0,0141	*	pTTAA	-0,1804	0,0432	*
pAA	-0,2150	0,0156	NS	pACCC	-0,1797	0,0441	NS
pGCC	0,2123	0,0170	NS	pCAA	-0,1796	0,0442	NS
pCGTG	0,2120	0,0172	*	pGG	0,1785	0,0455	NS
pAACG	0,2110	0,0177	NS	G+C%	0,1785	0,0456	NS
pCGAT	0,2077	0,0196	*	pCGTT	0,1774	0,0469	NS
pCGTC	0,2075	0,0197	*	...	...	...	...
pGGA	0,2048	0,0214	NS	pAGGG	0,0006	0,9946	NS
pATCG	0,2041	0,0219	NS	pTGA	-0,0003	0,9975	NS
pTGCC	0,2000	0,0248	*	pCACT	0,0003	0,9975	NS

[1]: os elementos genômicos (EG) que participaram da análise de correlações foram os seguintes: todas as sequencias de bi, tri e tetra nucleotídeos (pAA, PCC, pGG, pTT, pAC, pAG, pAT, ..., pTTTT); a porcentagem da soma das G e C por fragmento (G+C%); a freqüência dos motivos CCGNN; a intensidade de clivagem por radical OH^- (ClivOH); a quantidade de elementos MAR/SAR por fragmento (MAR cont); a freqüência de elementos MAR/SAR por kb em cada fragmento (MAR freq). A tabela foi truncada entre as linhas de ordem 35 e 338, pois, da linha 36 em diante, todas as correlações foram não significativas.

r: coeficiente de correlação de Pearson.

FDR: * e ** correspondem a valores significativos a 5% e 1%, respectivamente; NS corresponde a valores não significativos.

Além da variável MARcont, outras também apresentaram correlações significativas com ALD e $\boldsymbol{\rho}$, porém com coeficientes de correlação de sinais invertidos. A

freqüência dos tetra nucleotídeos CAAA e TAAG têm correlações positivas com ALD e negativas com $\boldsymbol{\rho}$. Já a do tetra nucleotídeo TGGA tem correlação negativa com ALD e positiva com $\boldsymbol{\rho}$.

Far-se-á algumas considerações sobre as frequências do tetra nucleotídeo TGGA e dos motivos CCGNN, que também apresentaram correlações positivas com a taxa de recombinação populacional ($\boldsymbol{\rho}$). Primeiramente, em relação ao motivo CCGNN, convém considerar que a formação e o posicionamento dos nucleossomos ao longo de uma molécula de DNA é dependente da sequência de nucleotídeos, sendo que, certos motivos de sequências favorecem a formação de nucleosssomos (Shrader & Crothers, 1989); enquanto outros propiciam a exclusão dos nucleossomos (Wang et al., 1996). Resultados do trabalho de Kirkpatrick et al. (1999) mostraram que repetições do motivo CCGNN elevam a taxa de recombinação meiótica em *Saccharomyces cerevisiae*. A detecção de correlação positiva entre a frequência do motivo CCGNN com as estimativas de $\boldsymbol{\rho}$ corroboram esses resultados, e sugere haver, de fato, uma associação positiva entre esse elemento genômico a as taxas de recombinação. O mecanismo que explicaria essa associação seria a determinação de um estado de maior acessibilidade à cromatina, promovido pelas regiões mais ricas em motivos CCGNN, pela sua propriedade de repelir a instalação de nucleossomos. Com menor densidade de nucleossomos, a cromatina ficaria mais propensa ao ataque das enzimas iniciadoras de DSB e, por conseguinte, à ocorrência de eventos de recombinação. Convém ressaltar que outros três tetranucleotídeos (CCGG, TCCG e GCCG), que contêm o trinucleotídeo CCG, também mostraram correlacões positivas e significativas com $\boldsymbol{\rho}$. Dentre vários pentanucleotídeos que apresentaram diferenças significativas de freqüências entre regiões *coldspot* e *hotspot* no genoma humano, Meyers et al. (2005) publicaram os cinco de maior magnitude nas diferenças. Desses cinco, três contêm o trinucleotídeo CCG. Kirkpatrick et al. (1999) identificaram que a repetição $(CCGAT)_{12}$, quando inserida em uma posição próxima ao gene HIS4, em *S. cerevisiae*, ativou, tanto a transcrição desse gene, quanto a recombinação meiótica na posição de inserção. Isso indica que variantes do motivo CCGNN podem ter funções ou efeitos distintos. Sugere-se, portanto, que em trabalhos futuros, se investigue as associações de cada membro da família CCGNN com a ocorrência de eventos de recombinação, e não a família como um todo.

Em segundo lugar, em relação à associação positiva entre o tetra nucleotídeo TGGA e $\boldsymbol{\rho}$, cita-se a pesquisa realizada por Cao et al. (1998). Os autores fizeram um

experimento *in vitro* utilizando milhares de sequências de DNA consituídas de 146 nucleotídeos (em ordem aleatória), flanqueados por adaptadores para amplificação por PCR. Essas sequências foram utilizadas para rescontituição de nucleossomos *in vitro*. Os fragmentos de DNA que não formavam nucleossomos eram purificados, amplificados por PCR e submetidos a ciclos subsequentes de reconstituição de nucleossomos. Após 17 ciclos de seleção, o material resultante estava rico em fragmentos que repeliam a formação de nucleossomos. Esses fragmentos, após clonagem e sequenciamento, se revelaram possuidores de longas repetições do motivo TGGA. A afinidade dessas repetições por octâmeros de histona era significativamente menor do que a metade da afinidade média dos outros fragmentos. Essa evidência pode ser somada à encontrada por Kirkpatrick et al. (1999), para configurar a situação na qual regiões com menor propensão à formação de nucleossomos são mais propícias à ocorrência de eventos de recombinação.

As variáveis pCAAA, pTAAG e MARcont mostraram-se correlacionadas positivamente com as médias das medidas do alcance do desequilíbrio de ligação, e negativamente com as médias de $\boldsymbol{\rho}$. Se regiões da classe *coldspots* são mais ricas em elementos MAR/SAR do que as regiões da classe *hotspots*, a consequência imediata seria pensar que este fator, de natureza estrutural, estaria moldando a distribuição dos eventos de recombinação ao longo do cromossomo. Porém, quando os valores de contagem desses elementos são ajustados para frequência de elementos MAR/SAR por kilo bases (variável MARfreq), a correlação se degenera e perde a significância estatística. Este fato pode ser explicado pela grande diferença entre os tamanhos médios dos fragmentos de cada classe: 1,9 kb para a classe dos *hotsposts* de recombinação e 17,2 kb para a classe dos *coldspots*. Quanto às variáveis pCAAA e pTAAG não foi encontrada nenhuma pesquisa anterior que tenha mencionado correlações entre estes dois tetra nucleotídeos e a ocorrência de *coldspots*.

A variável intensidade de clivagem por radical OH^- apresentou correlação positiva ($p < 0,02$) com as médias das estimativas de $\boldsymbol{\rho}$ (Tabela 6, décima linha). Essa variável traz informação de natureza estrutural, e segundo Greenbaum et al. (2007), proporciona uma informação associada à forma e estrutura local da dupla fita de DNA. Isto porque fornece uma medida da variação local do grau de acessibilidade dos solventes de DNA. Considerando-se que a correlação só se mostrou significativa entre a intensidade de clivagem por OH^- e as medidas de $\boldsymbol{\rho}$ e, considerando-se ainda a grande diferença entre os tamanhos dos fragmentos das duas classes analisadas, levanta-se a hipótese de que, na

escala mais ampla, o efeito da escala "escondeu" a provável correlação. Como os fragmentos da classe *hotspots* são, em média, bem menores, a correlação, nesta escala, apesar de tênue, se mostrou significativa estatisticamente ($p < 0,02$).

Diante das correlações significativas e de valores das estimativas do coeficiente de correlação relativamente altas, procedeu-se teste de médias para verificar se as freqüências de cada variável eram, de fato, distintas nas duas classes de fragmentos. Os resultados são aprsentados na Tabela 7.

Tabela 7. Comparação entre médias das freqüências de certos elementos genômicos nas classes de fragmentos *hotspot* e *coldspot*.

Variável	*Média Cold* [1]	*Média Hot* [2]	p-valor	Variável	*Média Cold* [1]	*Média Hot* [2]	p-valor
ClivOH	2,3659	0,2118	$< 10^{-6}$ **	pTCCG	0,5931	0,5226	0,0106 *
pAGTA	0,9906	0,8218	$< 10^{-6}$ **	pATCG	0,1565	0,1745	0,0117 *
pTAAG	0,4447	0,3848	$< 10^{-6}$ **	pCAA	0,2081	0,2611	0,0174 *
pCAAA	0,4389	0,5937	$< 10^{-6}$ **	pCGAT	0,2967	0,3339	0,0182 *
pTAGA	0,0933	0,1394	$< 10^{-6}$ **	pGGAA	0,4986	0,4093	0,0225 *
pGTAT	0,1372	0,1815	$< 10^{-6}$ **	pAACT	0,4960	0,5727	0,0251 *
pTTAG	0,2383	0,3183	$< 10^{-6}$ **	pGCCG	0,7743	0,8003	0,0265 *
pGTTA	0,1554	0,2056	$< 10^{-6}$ **	pTGCC	0,3672	0,3643	0,0288 *
pACCG	0,1625	0,1961	$< 10^{-6}$ **	pGCGT	0,3796	0,4167	0,0373 NS
pCTAG	1,3200	1,2405	$< 10^{-6}$ **	pGTA	0,0955	0,1196	0,0393 NS
pTTAA	2,2280	2,4034	$< 10^{-6}$ **	pTCTC	0,4260	0,4986	0,0531 NS
pAGAA	0,7893	0,7203	$< 10^{-6}$ **	pATGC	0,3329	0,3766	0,0659 NS
pGTCA	0,9093	0,8625	$< 10^{-6}$ **	pGGAT	0,3384	0,3461	0,0669 NS
pTAG	0,4777	0,4388	$< 10^{-5}$ **	pCTCT	0,3815	0,4761	0,0897 NS
pAAAG	0,9043	0,7938	$< 10^{-5}$ **	pTCTG	0,4710	0,4196	0,0982 NS
MARcont	0,4557	0,4269	$< 10^{-5}$ **	pCTAA	0,3818	0,3590	0,1118 NS
pGCAA	0,4069	0,3434	$< 10^{-5}$ **	pTCT	0,7877	0,6797	0,1172 NS
pTGGA	0,3988	0,3519	$< 10^{-4}$ **	pATAC	0,2444	0,2280	0,1360 NS
pTGC	0,2643	0,2401	0,0001 **	pTACA	0,5346	0,6204	0,1627 NS
pAGCT	0,3189	0,3162	0,0001 **	pTGCT	0,3780	0,4396	0,1698 NS
pGAGT	0,3672	0,3490	0,0005 **	pGATC	0,8488	0,9005	0,2034 NS
pTGG	0,4242	0,3926	0,0006 **	pGCTA	0,4308	0,4169	0,2910 NS
pCGTG	0,2654	0,2339	0,0030 **	pAAG	0,3736	0,4131	0,3541 NS
pCGT	0,3732	0,3269	0,0031 **	pTCTT	0,0092	0,0116	0,3556 NS
pCCGG	0,3255	0,3054	0,0038 **	pTAGT	0,2409	0,2615	0,4015 NS
pGTTG	0,4212	0,3786	0,0048 *	pTCCC	0,5967	0,7317	0,4634 NS
pGTGG	0,5007	0,4362	0,0057 *	pAAGC	1,3053	1,5732	0,4792 NS
pTCGT	0,4530	0,4064	0,0059 *	pGAAA	0,2076	0,2518	0,5295 NS
pCCGNN	2,3213	2,2497	0,0059 *	pGACT	0,1176	0,1666	0,5681 NS
pTCAA	2,5080	2,2799	0,0070 *	pTGAG	0,0810	0,1105	0,6796 NS
pAAGA	1,2149	1,1072	0,0077 *	pGTGT	0,2155	0,2724	0,8263 NS
pCGTC	1,0462	1,1204	0,0096 *	pGAAC	0,7642	0,6303	0,9174 NS

[1]: valores das médias aritméticas estimadas dentro da classe *coldspots*; [2]: valores das médias aritméticas estimadas dentro da classe *hotspots*; p-valor: esta coluna contém os p-valores seguidos de *, ** ou NS; * e ** indicam que as médias diferem entre si a 5% e 1% de probabilidade, respectivamente; NS corresponde a valores não significativos. Participaram dos testes de comparação de médias apenas as variáveis que apresentaram correlações significativas em pelo menos uma das duas classes de fragmentos (*hotspots* e *coldspots*).

Observando-se a Tabela 7 é possível notar que os sinais dos coeficientes de correlação de cada variável são invertidos da classe *hotspot* para a classe *coldspot*,

evidenciando o fato de que quando uma variável é positivamente correlacionada com uma das classes ela será negativamente correlacionada com a outra classe e vice-versa. Isso, somado ao fato de que as freqüências desses elementos genômicos são significativamente diferentes entre as duas classes, reforça a importância desses elementos na explicação das diferenças observadas na taxa de recombinação entre as duas classes de fragmentos. Poderia se recomendar o emprego destas variáveis como possíveis preditores da variação da taxa de recombinação, em análise de regressão múltipla ou *step wise*, por exemplo. Porém, antes disso, sugere-se que sejam investigadas as correlações destas variáveis com a taxa de recombinação ao longo do cromossomo 4 completo e não somente dentro das classes de fragmentos.

4.2.2 Análise de correlação entre a distribuição de elementos genômicos e as taxas de recombinação e de desequilíbrio de ligação ao longo do cromossomo, em várias escalas

Nessas novas análises as varáveis genômicas utilizadas foram: intensidade clivagem por radical OH^- (OH); percentagem de G+C (G+C%) e frequência de elementos MAR/SAR (MARfreq). Todas sistematizadas para os tamanhos de janelas de 6 a 20.000 locos. A Tabela 8 apresenta os resultados das análises de correlação entre as estimativas do desequilíbrio de ligação, em várias escalas, e as variáveis OH, G+C% e MARfreq. Todos os coeficientes de correlação de cada par de variáveis, em cada escala, apresentaram-se significativamente diferentes de zero. Existe uma flutuação em torno de zero para as escalas menores. À medida que se aumenta a escala, os valores absolutos dos coeficientes atingem um máximo, em determinada escala, e, em seguida voltam a diminuir. Como pode ser verificado, para as correlações dos três pares de variáveis (LD x OH, LD x G+C% e LD x MARfreq), o valor absoluto máximo do coeficiente de correlação é atingido na escala de 5.000 locos.

Tabela 8. Estimativas dos coeficientes de correlação de Pearson (*r*) entre as medidas do desequilíbrio de ligação (LD) e as frequências de alguns elementos genômicos[1] em várias escalas.

ESCALA[2]	LD x OH		LD x G+C%		LD x MARfreq	
(SNPs)	*r*	p-valor	*r*	p-valor	*r*	p-valor
2	0,0511	$<10^{-6}$	0,0470	$<10^{-6}$	-0,0119	0,015
6	0,0707	$<10^{-6}$	0,0591	$<10^{-6}$	-0,0258	$1,4*10^{-7}$
8	0,0713	$<10^{-6}$	0,0593	$<10^{-6}$	-0,0247	$<10^{-6}$

continua ...

Tabela 8. Continuação

ESCALA[2]	LD x OH		LD x G+C%		LD x MARfreq	
(SNPs)	*r*	p-valor	*r*	p-valor	*r*	p-valor
10	0,0686	$<10^{-6}$	0,0565	$<10^{-6}$	-0,0253	$<10^{-6}$
12	0,0646	$<10^{-6}$	0,0521	$<10^{-6}$	-0,0234	$1,7*10^{-6}$
16	0,0518	$<10^{-6}$	0,0387	$<10^{-6}$	-0,0165	0,001
20	0,0392	$<10^{-6}$	0,0269	$<10^{-6}$	-0,0092	0,061
50	-0,0291	$<10^{-6}$	-0,0327	$<10^{-6}$	0,0146	0,003
100	-0,1148	$<10^{-6}$	-0,1167	$<10^{-6}$	0,0537	$<10^{-6}$
200	-0,2036	$<10^{-6}$	-0,1999	$<10^{-6}$	0,0811	$<10^{-6}$
500	-0,3006	$<10^{-6}$	-0,2925	$<10^{-6}$	0,1587	$<10^{-6}$
1000	-0,3412	$<10^{-6}$	-0,3337	$<10^{-6}$	0,2508	$<10^{-6}$
2000	-0,3779	$<10^{-6}$	-0,3487	$<10^{-6}$	0,2645	$<10^{-6}$
5000	-0,4785	$<10^{-6}$	-0,4552	$<10^{-6}$	0,2970	$<10^{-6}$
10000	-0,3819	$<10^{-6}$	-0,3446	$<10^{-6}$	0,2141	$<10^{-6}$
20000	-0,4014	$<10^{-6}$	-0,3371	$<10^{-6}$	0,2227	$<10^{-6}$

[1]: LDxOH: correlação entre LD e a intensidade de clivagem por radical OH^-; LDxG+C%: correlação entre LD e a porcentagem da soma das bases G e C; LDxMARfreq: correlação entre LD e a frequência de elementos MAR/SAR;
[2]: Os números da coluna escala indicam a quantidade de locos SNPs que foram agregados numa janela de sistematização dos dados. Por exemplo, na escala 10 foram calculadas as médias das estimativas de 9 intervalos entre par de locos.

A Tabela 9 apresenta os resultados das análises de correlação entre as estimativas do alcance do desequilíbrio de ligação, em várias escalas, e as variáveis OH, G+C% e MARfreq. O mesmo padrão observado para as correlações apresentadas na Tabela 8 é, também aqui, observado, exceto pelo fato de que, neste caso, algumas correlações se mostraram não significativas. A escala na qual ocorreram os máximos dos valores absolutos dos coeficientes de correlação foi a de 10.000 locos.

Tabela 9. Estimativas dos coeficientes de correlação de Pearson (*r*) entre as medidas do alcance do desequilíbrio de ligação (ALD) e as frequências de alguns elementos genômicos[1] em várias escalas.

ESCALA[2]	ALD x OH		ALD x G+C%		ALD x MARfreq	
(SNPs)	*r*	p-valor	*r*	p-valor	*r*	p-valor
2	-0,0345	$<10^{-6}$	-0,0362	$<10^{-6}$	0,0085	0,081
6	-0,0106	0,030	-0,0089	0,069	-0,0087	0,076
8	-0,0002	0,961	0,0038	0,433	-0,0139	0,004
10	0,0075	0,126	0,0149	0,002	-0,0195	$6,9\ 10^{-5}$
12	0,0130	0,008	0,0230	$2,7*10^{-6}$	-0,0245	$<10^{-6}$
16	0,0219	$7,9\ 10^{-6}$	0,0353	$<10^{-6}$	-0,0316	$<10^{-6}$
20	0,0330	$<10^{-6}$	0,0481	$<10^{-6}$	-0,0385	$<10^{-6}$
50	0,0722	$<10^{-6}$	0,0967	$<10^{-6}$	-0,0885	$<10^{-6}$
100	0,1034	$<10^{-6}$	0,1378	$<10^{-6}$	-0,1477	$<10^{-6}$
200	0,1465	$<10^{-6}$	0,1803	$<10^{-6}$	-0,2151	$<10^{-6}$
500	0,2065	$<10^{-6}$	0,2476	$<10^{-6}$	-0,3126	$<10^{-6}$
1000	0,2690	$<10^{-6}$	0,3030	$<10^{-6}$	-0,3694	$<10^{-6}$

continua ...

Tabela 9. Continuação

ESCALA[2] (SNPs)	ALD x OH		ALD x G+C%		ALD x MARfreq	
	r	p-valor	*r*	p-valor	*r*	p-valor
2000	0,3359	$<10^{-6}$	0,3647	$<10^{-6}$	-0,4953	$<10^{-6}$
5000	0,3718	$<10^{-6}$	0,3795	$<10^{-6}$	-0,5922	$<10^{-6}$
10000	0,4379	$<10^{-6}$	0,4519	$<10^{-6}$	-0,6468	$<10^{-6}$
20000	0,3347	$<10^{-6}$	0,3987	$<10^{-6}$	-0,4974	$<10^{-6}$

[1]: ALDxOH: correlação entre ALD e a intensidade de clivagem por radical OH^-; ALDxG+C%: correlação entre ALD e a porcentagem da soma das bases G e C; ALDxMARfreq: correlação entre ALD e a frequência de elementos MAR/SAR;
[2]: Os números da coluna escala indicam a quantidade de locos SNPs que foram agregados numa janela de sistematização dos dados. Por exemplo, na escala 10 foram calculadas as médias das estimativas de 9 intervalos entre par de locos.

A Figura 34 ilustra o que se discutiu sobre o padrão observado nas Tabelas 8 e 9, em relação à variação das estimativas dos coeficientes de correlação ao longo de várias escalas. Mas, uma constatação que frusta expectativas iniciais é a de que as variáveis intensidade de clivagem por radical OH^- e a percentagem da soma das bases G e C, quando correlacionadas com as variáveis LD e ALD, propiciam exatamente a mesma informação. Observa-se que as linhas nas cores vermelha e sépia, que representam as correlações LD x OH e LD x G+C%, respectivamente, são paralelas entre si e quase coincidentes. O mesmo pode ser constatado para o par de linhas nas cores ciano e magenta, que representam as correlações ALD x OH e ALD x G+C%, respectivamente. Isso pode consistir num indício de que, se existe um mecanismo subjacente derivado da variação da porcentagem G+C ao longo do DNA que influencia as medidas do desequilíbrio de ligação, esse mecanismo é representado de forma bastante semelhante pelas variáveis G+C% e OH.

Havia a expectativa de que a variável intensidade de clivagem por radical OH^- carregasse uma informação nova, distinta daquelas baseadas em sequências de motivos ou conteúdo G+C. Sequências similares geram padrões de clivagem por radical OH^- similares. Porém, de acordo com Greenbaum et al. (2007), foram identificados longos fragmentos de DNA com baixíssima similaridade de sequência, que apresentaram padrões de clivagem por radical OH^- praticamente idênticos. Por outro lado, foram encontrados casos em que a mudança de apenas uma base no centro de duas sequências idênticas provocou uma drástica mudança na forma da curva de similaridade do padrão de clivagem por radical OH^-. Ainda, segundo esses mesmos autores, como o padrão de clivagem por OH^- é um reflexo da estrutura subjacente do DNA, os resultados dariam indícios de que sequências de DNA consideravalmente diferentes umas das outras poderiam compartilhar estrutura semelhante.

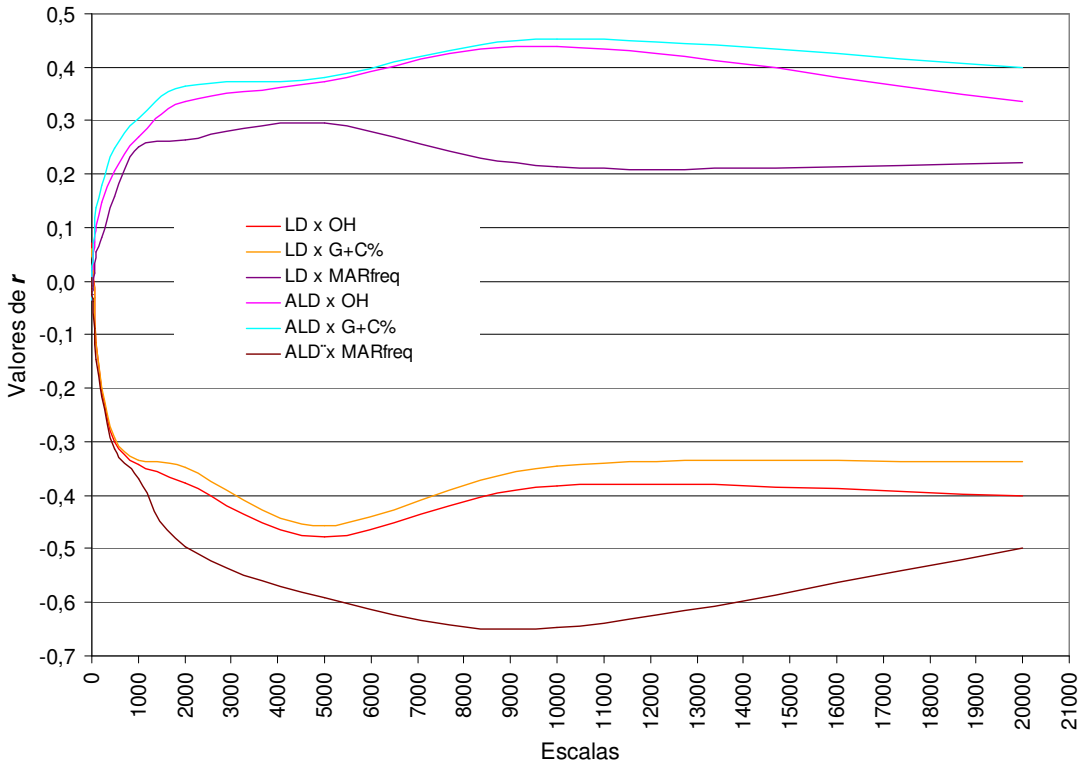


Figura 34. Curvas de variação das estimativas dos coeficientes de correlação (*r*) entre pares de variáveis ao longo de várias escalas. LD: desequilíbrio de ligação; ALD: alcance do desequilíbrio de ligação; OH: intensidade de clivagem por radical OH^-; G+C%: porcentagem da soma das bases G e C; MARfreq: freqüência de elementos MAR/SAR; Escalas: indica a quantidade de locos que foram agregados para calcular a média de cada variável numa dada janela. As janelas são sobrepostas e deslizantes, caminhando um loco a cada passo.

Greenbaum et al. (2007) investigaram essa possibilidade dividindo as sequências de DNA em *N*-nâmeros, variando de oito a 34 nucleotídeos de tamanho. Em seguida, para cada possível par de sequências, eles calcularam o grau de similaridade e os coeficientes de correlação de Pearson entre os padrões de clivagem por OH^-. Então, foi determinado a relação entre o grau de similaridade das sequências e as similaridades entre os padrões de clivagem por radical OH^-. As Figuras 35 e 36 mostram, respectivamente, a semelhança do padrão de clivagem por radical OH^- entre dois fragmentos de DNA, com 10% e 0% de similaridade de sequências.

Nas análises de correlação realizadas no escopo do presente trabalho utilizaram-se as médias das intensidades de clivagem por radical OH^- dentro do intervalo compreendido por uma janela. Esse detalhe da metodologia não capta a informação de natureza estrutural descoberta por Greenbaum et al. (2007). As médias das intensidades de clivagem representam as flutuações da porcentagem G+C em escalas mais amplas, pois, de fato, o conteúdo GC influencia o espectro de variação das medidas de várias propriedades

físicas e químicas do DNA.

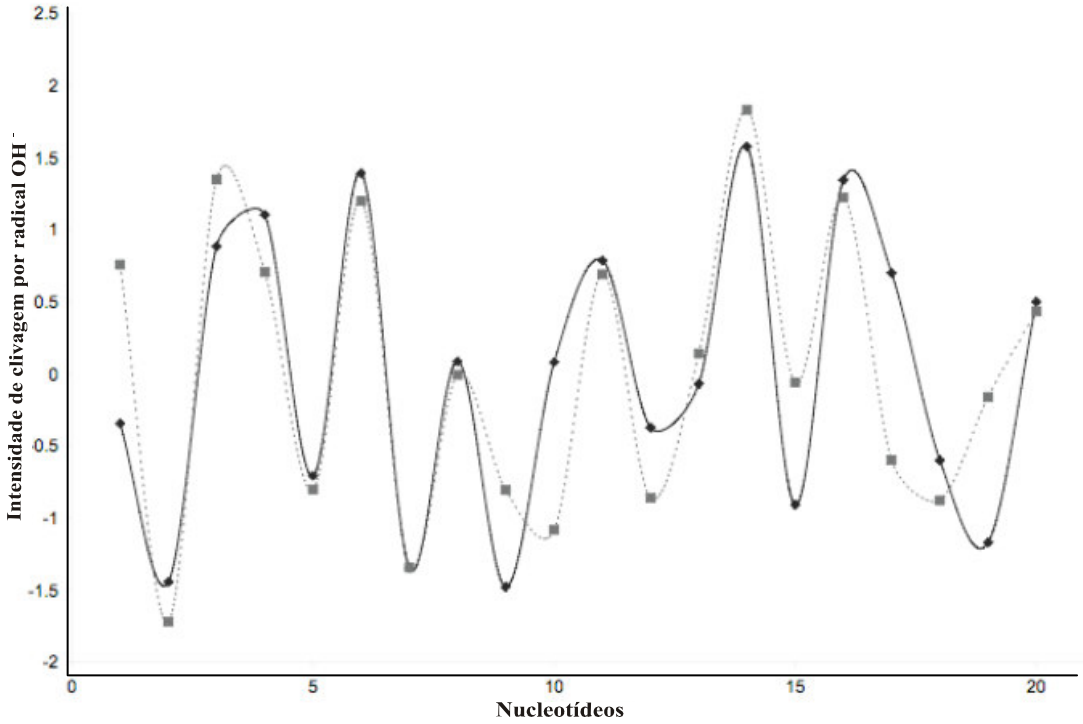


Figura 35. Curvas demonstrando o alto grau de similaridade do padrão de clivagem por radical OH^- (r = 0,81) para duas sequências de vinte nucleotídeos, com apenas 10% de similaridade na sequência. Fonte: Greenbaum et al. (2007).

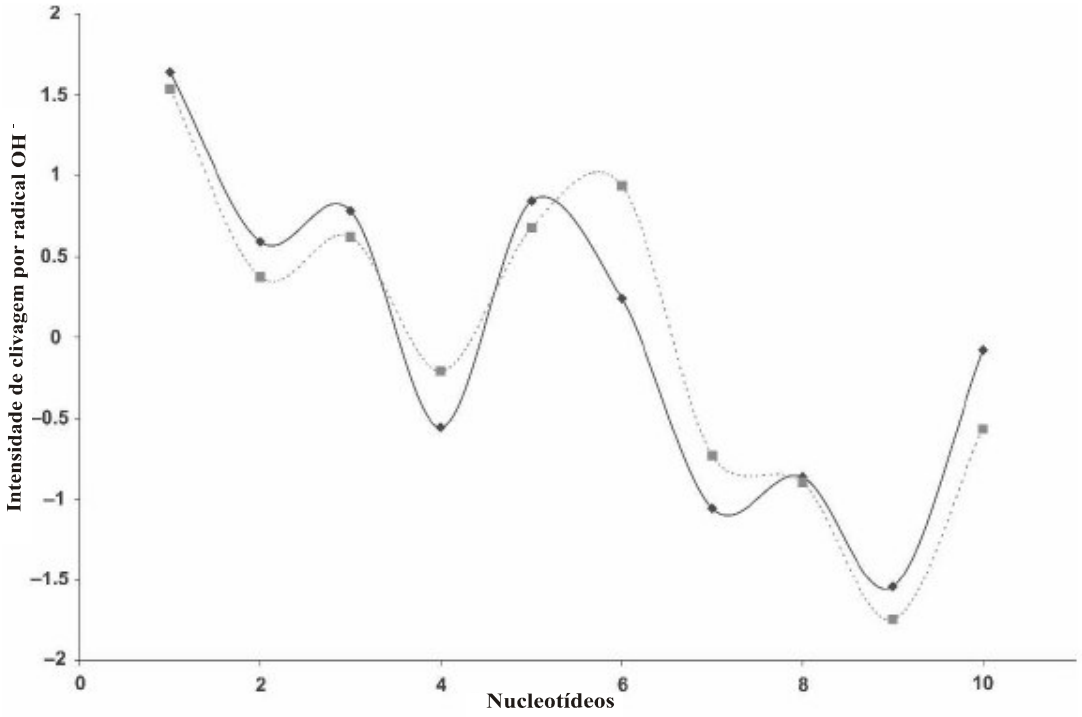


Figura 36. Curvas demonstrando o alto grau de similaridade do padrão de clivagem por radical OH^- (r = 0,94) para duas sequências de dez nucleotídeos com 0% de similaridade na sequência. Fonte: Greenbaum et al. (2007).

A sugestão para trabalhos futuros é que se proceda à busca de motivos estruturais, conforme metodologia utilizada por Greenbaum et al. (2007) e, uma vez identificados, proceder a contagem destes motivos dentro das janelas de cada escala, calculando-se sua frequências. Aí sim, buscam-se as possíveis correlações destes motivos com as variáveis associadas à recombinação e ao desequilíbrio de ligação.

Na Tabela 10 constam as estimativas dos coeficientes de correlação entre as estimativas da taxa de recombinação populacional (ρ) e as três variáveis utilizadas nas análises desta fase dos trabalhos. Para as três variáveis, como o número de escalas amostradas é muito pequeno, é possível observar apenas o comportamento inicial dos valores do coeficiente de correlação em torno de zero. Para a variável freqüência de elementos MAR/SAR (MARfreq), observa-se uma tendência inicial de diminuição dos valores dos coeficientes à medida em que aumenta a escala. Para as outras duas variáveis, intensidade de clivagem por radical OH^- (OH) e porcentagem da soma das bases G e C (G+C%), observa-se uma tendência de crescimento dos valores absolutos dos coeficientes de correlação. Nenhuma das três variáveis apresentaram sinais de máximos, nem mínimos para os valores dos coeficientes de correlação, dentro das escalas analisadas. Também fica evidente aqui que as variáveis OH e G+C% proporcionaram a mesma informação quando correlacionadas com ρ. Isto pode ser confirmado pelo grau de proximidade das curvas de cores ciano e azul da Figura 37.

Tabela 10. Estimativas dos coeficientes de correlação de Pearson (r) entre as medidas da taxa de recombinação (ρ) e as frequências de alguns elementos genômicos[1] em várias escalas.

ESCALA[2] (SNPs)	ρ x OH		ρ x G+C%		ρ x MARfreq	
	r	p-valor	r	p-valor	r	p-valor
8	0,0701	$<10^{-6}$	0,0816	$<10^{-6}$	-0,0049	0,319
10	0,0813	$<10^{-6}$	0,0862	$<10^{-6}$	-0,0072	0,143
12	0,0942	$<10^{-6}$	0,0941	$<10^{-6}$	-0,0120	0,014
16	0,1111	$<10^{-6}$	0,1012	$<10^{-6}$	-0,0193	$7{,}7*10^{-5}$
20	0,1186	$<10^{-6}$	0,1024	$<10^{-6}$	-0,0214	$1{,}3*10^{-5}$

[1]: ρ x OH: correlação entre ρ e a intensidade de clivagem por radical OH^-; ρ x G+C%: correlação entre ρ e a porcentagem da soma das bases G e C; ρ x MARfreq: correlação entre ρ e a frequência de elementos MAR/SAR;
[2]: Os números da coluna escala indicam a quantidade de locos SNPs que foram agregados numa janela de sistematização dos dados. Por exemplo, na escala 10 foram calculadas as médias das estimativas de 9 intervalos entre par de locos.

Constatada a existência de correlações significativas entre as variáveis analisadas, tomou-se, a título de exemplo, o caso da correlação entre o desequilíbrio de ligação e a intensidade de clivagem por radicais OH^-, cujo coeficiente de correlação de

Pearson tem estimativa no valor de r = -0,4785, na escala de 5.000 locos. Para este caso, elaborou-se o gráfico de dispersão entre as duas variáveis, apresentado na Figura 38.

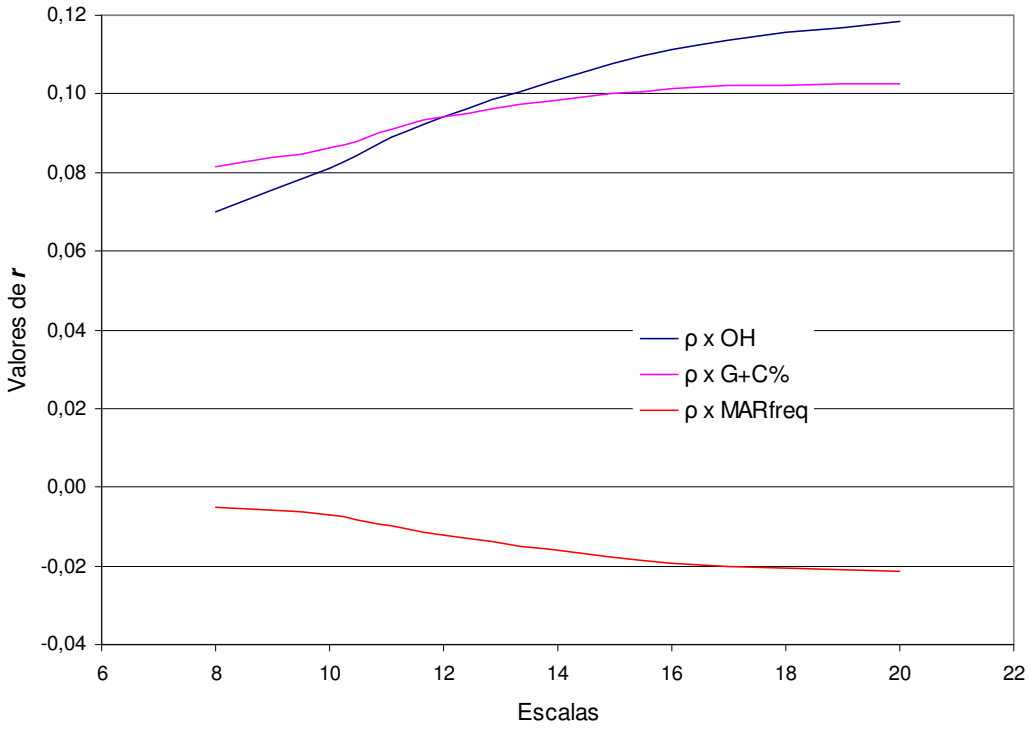


Figura 37. Curvas de variação das estimativas dos coeficientes de correlação (r) entre a taxa de recombinação populacional ($\boldsymbol{\rho}$) e algumas variáveis ao longo de várias escalas. OH: intensidade de clivagem por radical OH^-; G+C%: porcentagem da soma das bases G e C; MARfreq: freqüência de elementos MAR/SAR; Escalas: indica a quantidade de locos que foram agregados para calcular a média de cada variável numa dada janela. As janelas são sobrepostas e deslizantes, caminhando um loco a cada passo.

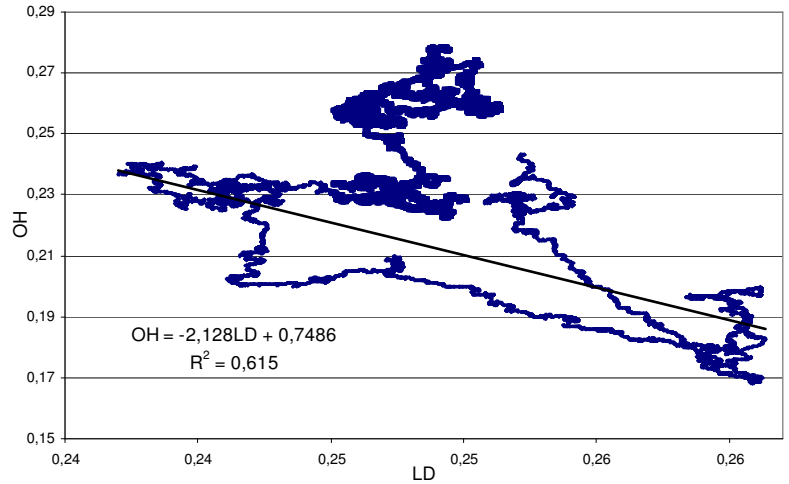


Figura 38. Gráfico da dispersão dos pontos correspondentes às variáveis desequilíbrio de ligação (LD) e intensidade de clivagem por radical OH^- (OH), usando estimativas calculadas na escala de 5.000 locos.

Diferentemente do que era esperado, os pontos não apresentaram um padrão de dispersão que se parecesse com uma nuvem de pontos espalhados numa região elipsoidal, cujo eixo maior seria representado pela reta da regressão linear simples. Ao contrário, ao se desenhar o gráfico percebeu-se que à medida que os pontos eram plotados uma trilha ia se configurando. Ficou explícita a presença de algum artefato de análise e/ou de sistematização dos dados. Esta constatação se deu para todos os demais gráficos de dispersão, de todos os outros pares de variáveis analisadas.

O diagnóstico levantado foi o de que o uso das janelas deslizantes sobrepostas gera uma autocorrelação acentuada entre os pontos. Por exemplo, no caso do uso de uma janela de 2.000 locos usam-se as estimativas dos 1.999 intervalos para o cálculo de uma média, que é ancorada em uma posição central no intervalo compreendido pelos 2000 locos. A estimativa do primeiro intervalo só participa do cálculo de uma média, a do segunto intervalo participa de duas e, assim por diante, de forma crescente até o milésimo intervalo, cuja estimativa participa de 1.000 médias. Do milésimo primeiro intervalo até o intervalo de número 2.000, as estimativas participam de 1.001 a 2.000 médias. Do intervalo 2.000 em diante todas as estimativas passam a participar de 2.000 médias. No outro extremo da série de dados, a forma de participação de cada estimativa é decrescente e simétrica à forma do início da série de dados.

Quanto maior o tamanho da janela utilizada maior será a intensidade da autocorrelação imposta às médias sistematizadas por janelas. Isso cria um artefato, pois o número de graus de liberdade é artificial. Por exemplo, numa série de 40 mil dados, usando-se uma janela deslizante sobreposta de 2.000 dados, é possível calcular 38 mil novas médias, cada uma calculada a partir de 2.000 valores; se a janela adotada tiver tamanho 20.000, é possível calcular 20.000 novas médias; então, o valor para os graus de liberdade seriam, respectivamente, 37.998 e 19.998. No caso do uso de janelas não sobrepostas, os valores dos graus de liberdade seriam, respectivamente, 18 e zero.

4.2.3 Avaliação do efeito de janelas deslizantes e sobrepostas nas correlações entre a distribuição de elementos genômicos e as taxas de recombinação e de desequilíbrio de ligação

Para observar o efeito do uso das janelas deslizantes sobrepostas sobre as estimativas dos coeficientes de correlação calculadas em várias escalas, realizou-se uma análise exatamente igual à realizada sobre as variáveis utilizadas no presente trabalho.

Porém, utilizando-se um conjunto de dados de duas variáveis x e y simuladas, geradas a partir de uma distribuição Normal, de média igual a 0,0 e variância igual a 1,0, de tal modo a garantir total independência entre as séries de dados das duas variáveis. Para cada variável foram gerados 41.760 dados, exatamente do mesmo tamanho do conjunto analisado no presente trabalho. O gráfico de dispersão entre as duas variáveis é apresentado na Figura 39.

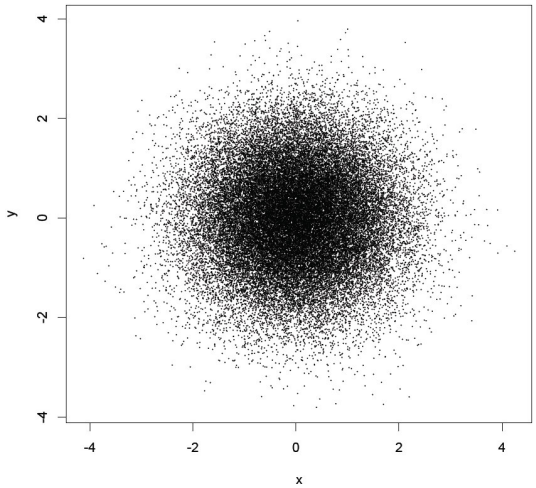


Figura 39. Gráfico da dispersão dos pontos correspondentes aos valores utilizados para a análise de correlação entre as variáveis x e y. A forma da nuvem de dispersão confirma a independência dos dados gerados.

A Figura 40 mostra o gráfico da variação dos valores do coeficiente de correlação ao longo de várias escalas, cujos tamanhos de janelas deslizantes sobrepostas variaram de dois a vinte mil dados agregados.

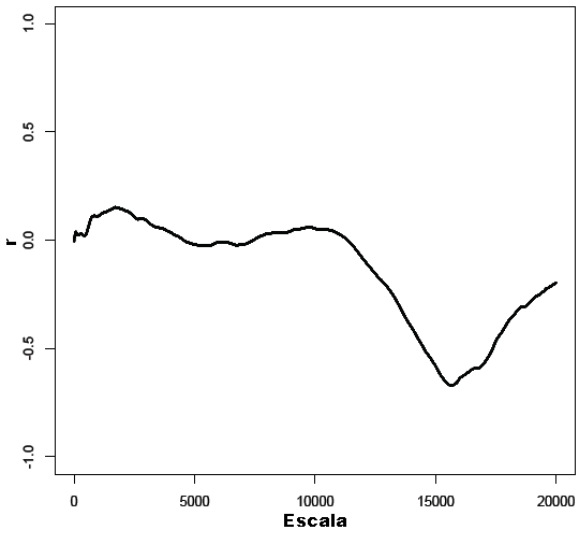


Figura 40. Gráfico da curva de variação das estimativas dos coeficientes de correlação (*r*) entre as variáveis x e y, ao longo de várias escalas, usando-se janelas deslizantes sobrepostas cujos tamanhos variaram de 2 a 20 mil dados agregados e o passo de caminhamento foi de um em um ponto na série de dados.

Essa simulação mostrou que a metodologia de sistematização das médias de valores de uma escala para outra, usando a técnica de janelas deslizantes sobrepostas, gera correlações artificiais importantes que não podem ser negligenciadas. No exemplo do conjunto de dados hipotéticos e aleatórios que foi utilizado, poderia-se chegar à conclusão errônea de que, na escala cuja janela tem tamanho de 15 mil pontos, a variável x tem correlação negativa forte com a variável y, com coeficiente de magnitude aproximadamente igual a -0,7. Outras simulações, usando conjuntos de dados contendo 5.000 valores normalmente distribuídos e janelas sobrespostas variando de um a 2.500 dados agregados, reforçaram a constatação de que o uso das janelas deslizantes sobrespostas proporciona o surgimento de correlações artificiais, conforme ilustrado na Figura 41.

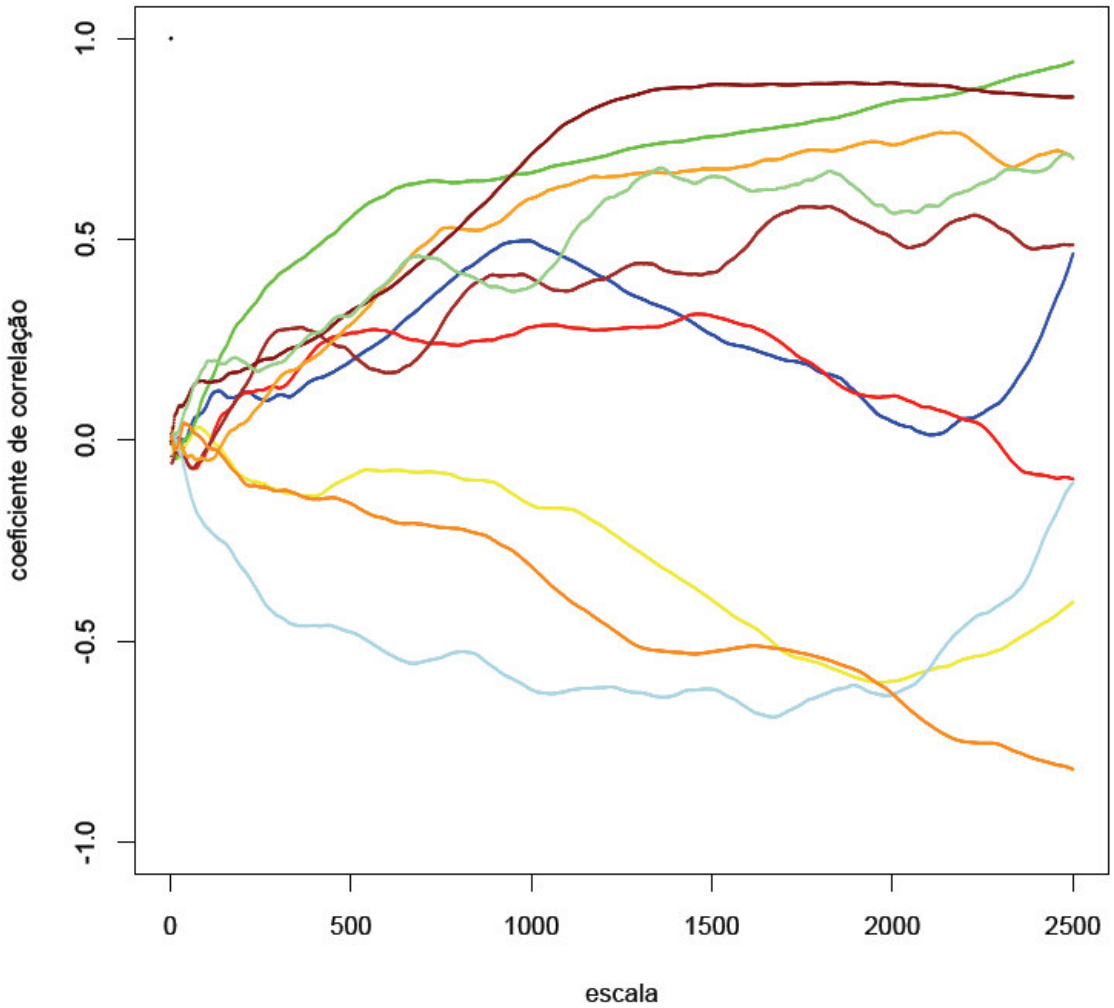


Figura 41. Gráfico da curva de variação das estimativas dos coeficientes de correlação (*r*) entre dois conjuntos de dados aleatórios, x e y, normalmente distribuídos, e, cada um, contendo 5 mil realizações. Utilizaram-se escalas correspondentes a janelas deslizantes sobrepostas cujos tamanhos variaram de 2 a 2.500 dados agregados e o passo de caminhamento foi de um em um ponto na série de dados.

Quando a simulação foi repetida sobre o mesmo conjunto de dados que gerou a Figura 40, usando-se janelas não sobrepostas, com tamanhos variando de dois a quatro mil dados, a configuração do gráfico passou a ser como o ilustrado na Figura 42.

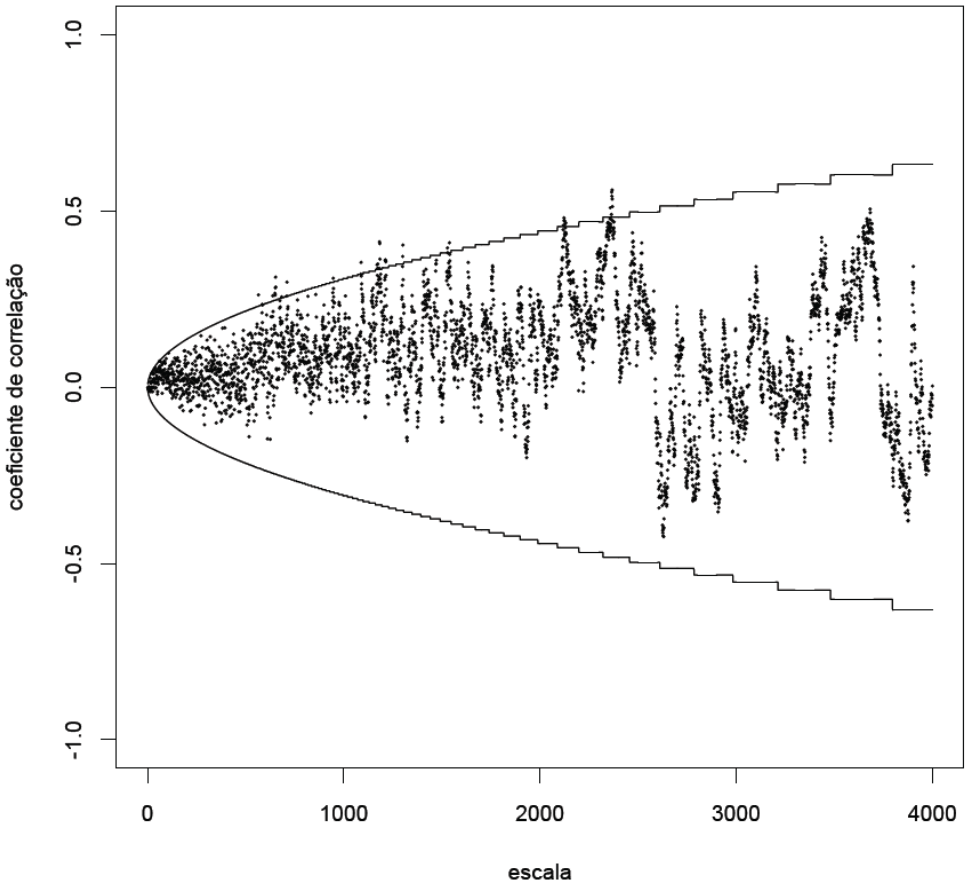


Figura 42. Gráfico da curva de variação das estimativas dos coeficientes de correlação (*r*) entre as variáveis x e y ao longo de várias escalas, usando-se janelas adjacentes não sobrepostas, cujos tamanhos variaram de dois a quatro mil dados agregados. Os pontos consistem na plotagem dos valores das estimativas de *r*. A linha, que lembra uma parábola com centro na origem dos eixos, representa os valores dos níveis de significância para rejeitar a hipótese de que *r* é igual a zero.

Observa-se que há uma oscilação dos valores das estimativas do coeficiente de correlação em torno de zero, com alguns valores absolutos próximos de 0,5. Porém, estimativas que apresentaram significância estatística são numericamente próximas às esperadas pela utilização de um nível de significância de 5%. Essa segunda simulação mostra que a técnica de sistematização baseada em janelas adjacentes não sobrepostas pode eliminar o artefato gerado pelo uso das janelas deslizantes sobrespostas.

Diante do exposto, cabe retomar alguns resultados do trabalho de Myers et al. (2005), realizado com dados do genoma humano, em que buscam modelos de predição da

recombinação com base em associações existentes entre a distribuição de elementos genômicos e a variação da recombinação ao longo de um genoma.

Myers et al. (2005) utilizaram 1,6 milhões de marcadores SNPs para genotipar amostras de três populações distintas e, em seguida, com base no modelo de coalescência, caracterizar a variação da recombinação ao longo do genoma humano. Após essa caracterização, os autores, usando determinado critério, identificaram regiões que foram classificadas como *hotspots* e outras como *cold sposts* de recombinação. Um subconjunto dos *hotspots*, cujos tamanhos dos fragmentos eram menores que 5 kb, foi utilizado para se realizar a busca de elementos genômicos que mostrassem associação com a variação da recombinação. Os autores escolheram cinco motivos repetitivos e 25 elementos genômicos, com as maiores diferenças significativas em suas quantidades entre os fragmentos *hotspots* e *coldspots* de recombinação, para elaborarem um modelo de predição da recombinação baseado em modelos lineares. Então, realizaram a sistematização de dados dos elementos preditores e da taxa de recombinação, nas escalas de 5 kb, 50 kb, 500 kb e 5 mb. A metodologia descrita no artigo não deixa claro se a técnica de sistematização utilizava janelas sobrepostas ou adjacentes. Os resultados informam que, para a escala de 5 mb, o modelo linear implementado explicava 42% da variação na taxa de recombinação. Para as escalas de 500 kb, 50 kb e 5 kb os modelos explicaram 34%, 15% e 4%, respectivamente.

Buscando-se um ponto de comparação para as escalas utilizadas nas análises do presente trabalho, a escala de 5 mb consistiria em apenas 3,7 janelas no cromossomo 4 de *A. thaliana*. Logo, no caso de janelas sobrepostas, a escala de 5 mb de Myers et al. (2005) corresponderia a uma janela de 11.000 SNPs para o presente trabalho. Ao se observar o valor de R^2 para a regressão linear simples, cujo gráfico é mostrado na Figura 38, apenas um elemento preditor, a intensidade de clivagem por radical OH^-, está explicando aproximadamente 61% da variação nas estimativas do desequilíbrio de ligação, na escala de 5.000 SNPs; a qual corresponderia a uma escala de 2,25 mb para os trabalhos de Myers et al. (2005).

A presença do efeito de janelas deslizantes e sobrepostas, identificado neste trabalho, gera falsas associações e/ou inflação das estimativas de correlação ao longo das escalas. Além disso, o que poderia ser considerado ainda pior, infla de modo diferenciado as correlações de uma escala em relação às de outras.

Outra comparação considerada importante para o contexto dessa discussão é a semelhança entre o gráfico da Figura 34, decorrente das análises de correlações ao longo

de várias escalas realizadas no presente estudo, e o gráfico da Figura 12, gerado por Kendal & Suomela (2005). Na Figura 12, os valores das estimativas do coeficiente de correlação chegam próximos a +0,9 e -0,7, na escala de 1.000 kb. Excluindo-se as escalas muito pequenas, que não aparecem no gráfico, todos os valores dos coeficientes de correlação são altamente significativos, conforme as delimitações das linhas centrais que representam o limite de significância naquela figura. Na metodologia descrita por Kendal & Suomela (2005) está explicitado, de forma clara, que as janelas utilizadas são adjacentes, não sobrepostas e com tamanhos variando de 10 kb a 1.000 kb. Os resultados por eles representados na Figura 12 consistem em médias calculadas nos cinco cromossomos de *A. thaliana*. Como é evidente a semelhança entre os dois gráficos, é provável que o diagnóstico para o caso das análises de Kendal & Suomela (2005) também seja o de que ocorreu um efeito resultante da sistematização dos dados. A explicação que se apresenta para esta situação é a seguinte: o uso de janelas sobrespostas provoca um aumento da autocorrelação entre os pontos de dados dentro da escala de sistematização; já o uso das janelas adjacentes, não sobrepostas, propaga um aumento de autocorrelação entre as escalas. De uma forma, ou de outra, ou por ambas, a influência desses efeitos se manifesta na geração de artefatos, ou seja, surgimento artificial de correlações fortes e altamente significativas.

Para eliminar o efeito das janelas sobrepostas, foi realizada uma nova sistematização dos dados para as variáveis utilizadas, utilizando-se a sistematização por janelas adjacentes, não sobrepostas. Para facilidades computacionais foram utilizadas 21 escalas de janelas com tamanhos sub-múltiplos de 41.760, que é o número total de intervalos entre marcas no cromossomo 4 de *A. thaliana*. Os tamanhos de janelas utilizados foram: 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 16, 18, 20, 30, 40, 48, 96, 144, 288, 522 e 1.044 locos. Os graus de liberdade variaram, nesta ordem, de 20.880 a 40. Como dentro de cada escala realizaram-se 21 testes de t para verificar a significância estatística dos coeficientes de correlação calculados, foi aplicado o procedimento FDR para cada par de variáveis analisadas. A Tabela 11 apresenta os resultados das análises de correlação após esta nova sistematização dos dados.

As correlações entre as estimativas do desequilíbrio de ligação e as frequências de elementos MAR/SAR que se apresentaram como significativas na análise com janelas sobrepostas, já não o são na análise que utiliza janelas adjacentes não sobrepostas. Os valores dos coeficientes de correlação oscilam em torno de zero, chegam a um pico

negativo na escala correspondente a janela de 30 SNPs e, em seguida se estabilizam numa tendência crescente até a escala correspondente a janela de 1.044 SNPs.

Tabela 11. Estimativas dos coeficientes de correlação de Pearson (*r*) entre as medidas do desequilíbrio de ligação (LD) e as frequências de alguns elementos genômicos em várias escalas.

Variáveis[1]	*r*	t	Janelas[2]		GL	p-valor	FDR
			Tam.	Qtd.			
LD x MARfreq	-0,0144	-2,0854	2	20880	20878	0,01852	NS
LD x MARfreq	0,0025	0,2891	3	13920	13918	0,38627	NS
LD x MARfreq	0,0033	0,3350	4	10440	10438	0,36880	NS
LD x MARfreq	0,0037	0,3405	5	8352	8350	0,36676	NS
LD x MARfreq	0,0040	0,3340	6	6960	6958	0,36919	NS
LD x MARfreq	-0,0078	-0,5632	8	5220	5218	0,28667	NS
LD x MARfreq	-0,0078	-0,5291	9	4640	4638	0,29840	NS
LD x MARfreq	-0,0093	-0,5982	10	4176	4174	0,27488	NS
LD x MARfreq	-0,0136	-0,8031	12	3480	3478	0,21097	NS
LD x MARfreq	-0,0041	-0,2189	15	2784	2782	0,41339	NS
LD x MARfreq	-0,0100	-0,5092	16	2610	2608	0,30534	NS
LD x MARfreq	-0,0155	-0,7452	18	2320	2318	0,22812	NS
LD x MARfreq	0,0017	0,0789	20	2088	2086	0,46856	NS
LD x MARfreq	-0,0157	-0,5850	30	1392	1390	0,27933	NS
LD x MARfreq	-0,0003	-0,0087	40	1044	1042	0,49654	NS
LD x MARfreq	0,0182	0,5371	48	870	868	0,29567	NS
LD x MARfreq	0,0690	1,4399	96	435	433	0,07531	NS
LD x MARfreq	0,0722	1,2293	144	290	288	0,10998	NS
LD x MARfreq	0,0383	0,4586	288	145	143	0,32362	NS
LD x MARfreq	0,0959	0,8505	522	80	78	0,19882	NS
LD x MARfreq	0,1550	0,9674	1044	40	38	0,16973	NS
LD x OH	0,0617	8,9383	2	20880	20878	0,00000	**
LD x OH	-0,0204	-2,4082	3	13920	13918	0,00802	*
LD x OH	-0,0225	-2,3015	4	10440	10438	0,01069	*
LD x OH	-0,0124	-1,1295	5	8352	8350	0,12935	NS
LD x OH	-0,0057	-0,4786	6	6960	6958	0,31613	NS
LD x OH	0,0662	4,7923	8	5220	5218	0,00000	**
LD x OH	0,0165	1,1237	9	4640	4638	0,13059	NS
LD x OH	0,0154	0,9927	10	4176	4174	0,16045	NS
LD x OH	0,0665	3,9288	12	3480	3478	0,00004	**
LD x OH	0,0458	2,4202	15	2784	2782	0,00779	*
LD x OH	0,0078	0,3999	16	2610	2608	0,34465	NS
LD x OH	0,0455	2,1914	18	2320	2318	0,01426	*
LD x OH	0,0297	1,3567	20	2088	2086	0,08751	NS
LD x OH	0,0182	0,6794	30	1392	1390	0,24851	NS
LD x OH	0,0255	0,8226	40	1044	1042	0,20546	NS
LD x OH	-0,0353	-1,0414	48	870	868	0,14899	NS
LD x OH	0,0032	0,0669	96	435	433	0,47337	NS
LD x OH	-0,0420	-0,7132	144	290	288	0,23815	NS
LD x OH	-0,2315	-2,8454	288	145	143	0,00254	*
LD x OH	-0,2556	-2,3354	522	80	78	0,01105	*
LD x OH	-0,2433	-1,5460	1044	40	38	0,06519	NS
LD x G+C%	0,0498	7,2087	2	20880	20878	0,00000	**
LD x G+C%	-0,0238	-2,8095	3	13920	13918	0,00248	*
LD x G+C%	-0,0238	-2,4359	4	10440	10438	0,00744	*
LD x G+C%	-0,0151	-1,3801	5	8352	8350	0,08379	NS
LD x G+C%	-0,0097	-0,8131	6	6960	6958	0,20810	NS

continua ...

Tabela 11. Continuação

Variáveis[1]	*r*	t	Janelas[2]		GL	p-valor	FDR
			Tam.	Qtd.			
LD x G+C%	0,0547	3,9554	8	5220	5218	0,00004	**
LD x G+C%	0,0189	1,2866	9	4640	4638	0,09915	NS
LD x G+C%	0,0181	1,1696	10	4176	4174	0,12112	NS
LD x G+C%	0,0487	2,8734	12	3480	3478	0,00204	*
LD x G+C%	0,0323	1,7068	15	2784	2782	0,04399	NS
LD x G+C%	0,0116	0,5910	16	2610	2608	0,27730	NS
LD x G+C%	0,0345	1,6634	18	2320	2318	0,04818	NS
LD x G+C%	0,0134	0,6115	20	2088	2086	0,27048	NS
LD x G+C%	0,0188	0,7010	30	1392	1390	0,24172	NS
LD x G+C%	0,0172	0,5555	40	1044	1042	0,28933	NS
LD x G+C%	-0,0385	-1,1348	48	870	868	0,12840	NS
LD x G+C%	-0,0074	-0,1547	96	435	433	0,43855	NS
LD x G+C%	-0,0535	-0,9089	144	290	288	0,18208	NS
LD x G+C%	-0,2311	-2,8409	288	145	143	0,00258	*
LD x G+C%	-0,2548	-2,3267	522	80	78	0,01129	*
LD x G+C%	-0,2812	-1,8064	1044	40	38	0,03939	NS

[1]: MARfreq: freqüência de elementos MAR/SAR; OH: intensidade de clivagem por radical OH^-; G+C%: porcentagem da soma das bases G e C.
[2]: Tam.: tamanho da janela utilizada em número de locos; Qtd.: quantidade de janelas adjacentes.
GL: graus de liberdade associado ao teste t. FDR: * e ** indicam que as médias diferem entre si a 5% e 1% de probabilidade, respectivamente; NS corresponde a valores não significativos.

Nesta nova análise, as variáveis clivagem por radical OH^- (OH) e porcentagem da soma das bases G e C (G+C%) continuam sendo redundantes no que concerne suas relações com a variável desequilíbrio de ligação. Ocorre, neste caso, coincidência quase completa, desde as escalas nas quais os coeficientes de correlação se apresentam com significância estatística, até a magnitude e o sinal dos coeficientes de correlação. O fato de duas variáveis apresentarem associação estatisticamente significativa em mais de uma escala significa, portanto, que os sinais que descrevem a variação dessas variáveis ao longo do cromossomo podem compartilhar algumas características de periodicidade e frequência.

Para ilustrar a remoção da autocorrelação identificada nas análises com janelas deslizantes e sobrepostas, escolheu-se plotar o gráfico de dispersão para o par de variáveis LD x OH, numa escala cuja janela tem tamanho de 522 locos e que resultou num coeficiente de correlação igual a -0,2556 (Figura 43). Conforme pode-se observar neste gráfico, a nuvem de pontos já não se apresenta na forma de uma trilha sequencial de pontos, como na Figura 38.

Apesar de ter sido encontrada uma associação entre as estimativas do desequilíbrio de ligação e as variáveis genômicas G+C% e OH, os resultados apresentados por Kendal & Suomela (2005), em particular a forma das curvas nos gráficos da Figura 12, apontam para a existência de algum artefato provocado pela propagação da autocorrelação dos dados de uma escala para outra. Para conferir a forma adquirida pelas curvas das

estimativas dos coeficientes de correlação, ao longo de várias escalas, após a adoção das janelas adjacentes não sobrepostas, confeccioou-se o gráfico da Figura 44.

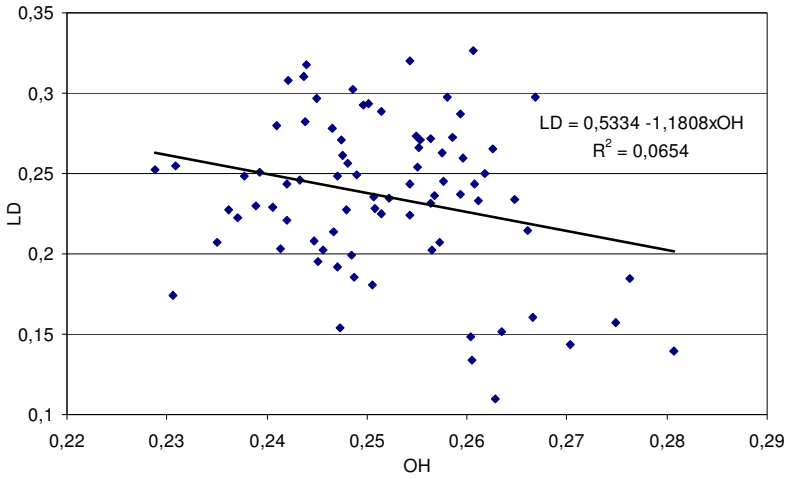


Figura 43. Gráfico da dispersão dos pontos correspondentes aos valores utilizados para a análise de correlação entre as variáveis desequilíbrio de ligação (LD) e intensidade de clivagem por radical OH^- (OH), usando janelas adjacentes, não sobrepostas de tamanho igual a 522 locos.

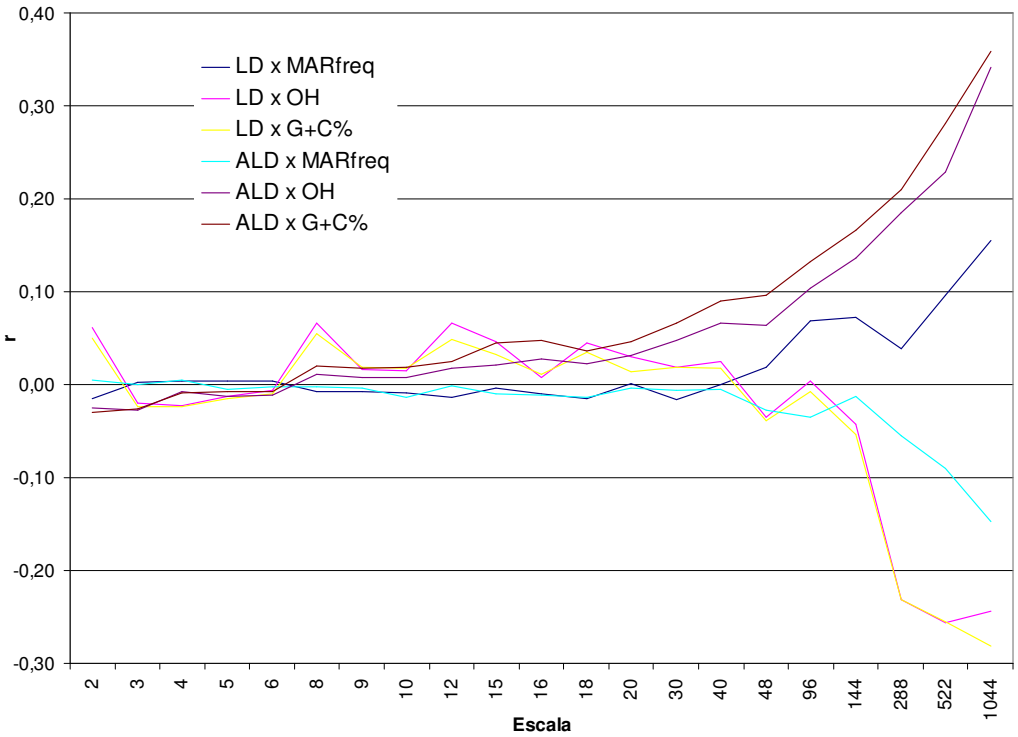


Figura 44. Curvas de variação das estimativas dos coeficientes de correlação (*r*) entre pares de variáveis ao longo de várias escalas. LD: desequilíbrio de ligação; ALD: alcance do desequilíbrio de ligação; OH: intensidade de clivagem por radical OH^-; G+C%: porcentagem da soma das bases G e C; MARfreq: freqüência de elementos MAR/SAR; Escalas: indica a quantidade de locos que foram agregados para calcular a média de cada variável numa dada janela. As janelas são adjacentes e não sobrepostas.

Nas escalas de janelas até 96 locos, o comportamento das linhas é semelhante ao apresentado pela nuvem de pontos da Figura 42, caracterizado por uma oscilação em torno de zero e pela não significância estatística das estimativas. De certo ponto em diante, neste caso da escala 96 locos em diante, os valores das estimativas dos coeficientes de correlação parecem começar a sofrer o efeito da autocorrelação propagada de uma escala para outra. Seria o mesmo que dizer que uma escala não é totalmente idependente da outra. Caso não houvesse uma limitação dos graus de liberdade, as análises poderiam ter continuado em escalas maiores e, provavelmente seria obtido um gráfico semelhante ao da Figura 12, elaborado por Kendal & Suomela (2005). É importante ressaltar que a janela de 1044 locos da Figura 44, corresponde a 470 kb da Figura 12, ponto no qual as estimativas de correlação média para todo o genoma de *A. thaliana* atingem certa estabilização. Daí a suposição de que, logo após o crescimento vertiginoso das estimativas do coeficiente de correlação, que se dá entre as escalas 96 e 1.044, venha a ocorrer nova estabilização.

Como foi verificado, o uso de janelas deslizantes sobrepostas pode acarretar artefatos para a análise de correlações entre elementos genômicos. Portanto, faz-se necessário lançar mão de outras técnicas e/ou métodos de análise livres desse incômodo. Audit et al. (2002) informam que a análise *wavelet* decompõe a variação total presente em uma série de dados (temporal ou espacial, como ao longo de uma sequência) em seus componenetes ao longo de várias escalas. Para cada escala são calculados coeficientes que captam a variação dentro de uma dada escala. A técnica de estimação dos coeficientes garante a obtenção de vetores ortogonais entre si, ou seja, os coeficientes de uma escala são independentes dos coeficientes das outras escalas. Os coeficientes de cada escala podem ser considerados como os parâmetros de ondas senoidais. A superposição de um sinal senoidal de menor periodicidade, ou maior frequência, sobre outro sinal de menor frequência gera um terceiro sinal diferente dos sinais parentais. Com base nessa técnica, um sinal pode ser totalmente reconstruído se os coeficientes das várias escalas forem fornecidos. Esse princípio tem sido utilizado para a compactação e transmissão de vários tipos de dados.

Portanto, como sugere Lió (2003), a busca por prováveis associações entre elementos genômicos pode ter mais sucesso se forem utilizados os coeficientes de cada escala e não as medidas sistematizadas das próprias variáveis ao longo de escalas. Os coeficientes de cada escala captariam, de forma mais adequada, o padrão presente nos sinais, diminuindo os ruídos e, portanto, modelando mais adequadamente a variação em

cada escala.

Modelos que se propõem a predizer taxas de recombinação ao longo de um cromossomo poderiam, então, ser baseados em multi-escalas. Isto é, para cada escala poderia ser realizada uma regressão linear *step wise* identificando as variáveis que melhor explicam a variação da recombinação naquela escala. Após realizadas estas regressões, o modelo geral se constituiria de uma soma das predições em todas as escalas. A aditividade entre escalas estaria garantida pela estimação de coeficientes idependentes entre as escalas como as proporcionadas pela técnica de análise *wavelet*.

Sugere-se, portanto, que estudos futuros devam explorar mais detalhadamente os resultados apresentados nas Tabelas 5 e 6, das correlações de algumas sequências com a ocorrência de eventos de recombinação e com o alcance do desequilíbrio de ligação. Porém, numa estrutura de dados orientada pelas técnicas da análise *wavelet*.

5 CONCLUSÕES

Os resultados dos estudos desenvolvidos no presente trabalho permitem concluir:

1. Ao longo do cromossomo 4 de *A. thaliana*, os eventos de recombinação ocorrem de forma concentrada, pois proporções de 50% a 60% dos eventos de recombinação ocorrem em apenas 13% a 20% da sequência de DNA.
2. A contagem de elementos MAR/SAR, a média da intensidade de clivagem por radical OH^- e as frequências do motivo CCGNN, de cinco tri nucleotídeos (CAA, CGT, TAG, TGC e TGG) e de 32 tetra nucleotídeos (AAAG, AACT, AAGA, ACCG, AGAA, AGCT, AGTA, ATCG, CAAA, CCGG, CGAT, CGTC, CGTG, CTAG, GAGT, GCAA, GCCG, GGAA, GTAT, GTCA, GTGG, GTTA, GTTG, TAAG, TAGA, TCAA, TCCG, TCGT, TGCC, TGGA, TTAA e TTAG) apresentaram diferenças significativas entre as classes de fragmentos *hotspots* e *coldspots*.
3. Além disso, todos esses elementos genômicos apresentaram correlação significativa com a taxa de recombinação populacional ($\boldsymbol{\rho}$) e com o alcance do desequilíbrio de ligação (ALD), sendo que, seus respectivos coeficientes de correlação tem sinal invertido com $\boldsymbol{\rho}$, em relação ao sinal com ALD.
4. A variável genômica porcentagem da soma das bases G e C (G+C%) tem associações importantes com a variação do desequilíbrio de ligação ao longo do cromossomo 4 de *A. thaliana*, e estas associações ocorrem em mais de uma escala.
5. A variável genômica média da intensidade de clivagem por radical OH^-, nas análises de correlações com as estimativas de desequilíbrio de ligação, proporciona informação redundante com a variável G+C%.
6. O uso de janelas deslizantes sobrepostas gera distorções importantes nas análises de correlações entre as distribuições dos elementos genômicos e a variação de taxas de recombinação e de desequilíbrio de ligação, provocando, artificialmente, correlações supostamente fortes e significativas em uma ou mais escalas.
7. A intensidade relativa e o posicionamento de algumas feições na curva da variação da taxa de recombinação ao longo do cromossomo 4 de *A. thaliana* parecem ter

razoável repetibilidade, pois são detectadas por meio do emprego de métodos e populações distintas.

8. Considerando o critério utilizado para a detecção de *hotspots*, os fragmentos de DNA sobre os quais recaem as estimativas mais extremas de taxas de recombinação se caracterizam por terem tamanho médio menor que 2 kb e por terem uma distribuição concentrada, preferencialmente, na região correspondente ao terço médio do cromossomo 4 de *A. thaliana*.
9. Considerando o critério utilizado para a detecção de *coldspots*, os fragmentos de DNA sobre os quais recaem as estimativas mais extremas do alcance do desequilíbrio de ligação se caracterizam por terem tamanho médio maior que 17 kb e distribuição concentrada, preferencialmente, nas extremidades distais do cromossomo 4 de *A. thaliana*.

6 REFERÊNCIAS

AKHUNOV, E. D.; GOODYEAR, A. W.; GENG, S.; QI, L.; ECHALIER, B.; GILL, B. S.; GUSTAFSON, J. P.; LAZO, G.; CHAO, S.; ANDERSON, O. D.; LINKIEWICZ, A. M.; DUBCOVSKY, J.; LA ROTA, M.; SORRELLS, M. E.; ZHANG, D.; NGUYEN, H.; KALAVACHARLA, V.; HOSSAIN, K.; KIANIAN, S. F.; PENG, J.; LAPITAN, N. L. V.; GONZALES-HERNANDEZ, J. L.; ANDERSON, J. A.; CHOI, D. W.; CLOSE, T. J.; DILBIRLIGI, M.; GILL, K. S.; WALKER-SIMONS, M. K.; STEBER, C.; McGUIRE, P. E.; QUALSET, C. O.; DVORAK, J. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. **Genome Research**, Cold Spring Harbor, v. 13, p. 753-763, 2003.

ALLARD, R. W. Evidence for genetic restriction of recombination in the lima bean. **Genetics**, Bethesda, v. 48, p. 1389-1395, 1963.

ANDERSON, L. K.; HOOKER, K. D.; STACK, S. M. The distribution of early recombination nodules on zygotene bivalents from plants. **Genetics**, Bethesda, v. 159, p. 1259-1269, 2001.

ANDERSON, L. K.; DOYLE, G. G.; BRIGHAM, B.; CARTER, J.; HOOKER, K.D.; LAI, A.; RICE, M.; STACK, S. M. High-resolution croosover maps for each bivalent of Zea mays using recombination nodules. **Genetics**, Bethesda, v. 165, p. 849-865, 2003.

ANDERSON, L. K.; STACK, S. M. Recombination nodules in plants. **Cytogenetic and Genome Research**, Basel, v. 109, n. 1-3, p. 198-204, 2005.

ANDERSON, L. K.; LAI, A.; STACK, S. M.; RIZZON, C.; GAUT, B. S. Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. **Genome Research**, Cold Spring Harbor, v. 16, p. 115-122, 2006.

AUDIT, B.; BACRY, E.; MUZY, J. F.; ARNEODO, A. Wavelet-based estimators of scaling behavior. **IEEE Transactions on information theory**, New York, v. 48, n. 11, p. 2938-2954, 2002.

AUTON, A. The estimation of recombination rates from population genetic data. 2007. 202 f. **Thesis** (Doctor of Philosophy) Hertford College, University of Oxford, Trinity, 2007.

AUTON. A.; MCVEAN, G. A. T. Recombination rate estimation in the presence of hotspots. **Genome research**, Cold Spring Harbor, v. 17, n. 8, p. 1219-1227, 2007.

BAGSHAW, A. T. M.; PITT, J. P. W.; GEMMELL, N. J. Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots. **BMC Genomics**, London, v. 7, n. 179, 2006.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a pratical and powerful approach to multiple testing. **Journal of the Royal Statistical Society, Series B (Statistical Methodology)**, Hoboken, v. 57, n. 1, p. 289-300, 1995.

BENNETZEN, J. L.; MA, J.; DEVOS, K. M. Mechanisms of recent genome size variation in flowering plants. **Annals of Botany**, Oxford, v. 95, p. 127-132, 2005.

BISHOP, D.; ZICKLER, D. Early decision meiotic crossover interference prior to stable strand exchange and synapsis. **Cell**, Cambridge, v. 117, n. 1, p. 9-15, 2004.

BLUMENTAL-PERRY, A.; ZENVIRTH, D.; KLEIN, S.; ONN, I.; SIMCHEN, G. DNA motif associated with meiotic double-strand break regions in Saccharomyces cerevisiae. **EMBO Reports**, Cambridge, v. 1, n. 3, p. 232-238, 2000.

BOREVITZ, J. O.; NORDBORG, M.; MARJORAM, P.; ZOELLNER, S. **Genome wide association mapping in Arabidopsis thaliana**. Chigago: University of Chigago/ Department of Ecology and Evolution, 2009a. Disponível em <http://naturalsystems.uchicago.edu/naturalvariation/hapmap/ distFinal360Name.pdf>. Acesso em: 10 abr. 2010.

BOREVITZ, J. O.; NORDBORG, M.; MARJORAM, P.; ZOELLNER, S. **Genome wide association mapping in Arabidopsis thaliana**. Chigago: University of Chigago/ Department of Ecology and Evolution, 2009b. Disponível em <http://naturalsystems.uchicago.edu/naturalvariation/hapmap/NIHyear2.pdf>. Acesso em: 10 abr. 2010.

BOWERS, J. E.; ARIAS, M. A.; ASHER, R.; AVISE, J. A.; BALL, R. T.; BREWER, G. A.; BUSS, R. W.; CHEN, A. H.; EDWARDS, T. M.; ESTILL, J. C.; EXUM, H. E.; GOFF, V. H.; HERRICK, K. L.; STEELE, C. L. J.; KARUNAKARAN, S.; LAFAYETTE, G. K.; LEMKE, C.; MARLER, B. S.; MASTERS, S. L.; McMILLAN, J. M.; NELSON, L. K.; NEWSOMW, G. A.; NWAKANMA. C. C.; ODEH, R. N.; PHELPS, C. A.; RARICK, E. A.; ROGERS, C. J.; RYAN, S. P.; SLAUGHTER, K. A.; SODERLUND, C. A.; TANG, H.; WING, R. A.; PATERSON, A. H. Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. **Proceedings of the National Academy of Science of the USA**, Washington, v. 102, p. 13206-13211, 2005.

BREYNE, P; MONTAGU, M. V.; GHEYSEN, G. The role of scaffold attachment regions in the structural and functional organization of plant chromatin. **Transgenic Research**, Vienna, v. 3, n. 3, p. 195-202, 1994.

CAO, H.; WIDLUND, H. R.; SIMONSON, T.; KUBISTA, M. TGGA repeats impair nucleosome formation. **Journal of Molecular Biology**, Maryland, v. 281, n. 2, p. 253-260, 1998.

CARNEGIE INSTITUTION FOR SCIENCE. **The Arabidopsis Information Resource (TAIR)**. Stanford: CIFS/ Department of Plant Biology, 2009. Disponível em <ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/chr4.fas>. Acesso em: 10 jul. 2009.

CHEN, L.; ZHAO, H. Negative correlation between compositional symmetries and local recombination rates. **Bioinformatics**, Oxford, v. 21, n. 21, p. 3951-3958, 2005.

CHOO, H. H. A. Why is the centromere so cold? **Genome Research**, Cold Spring Harbor, v. 8, p. 81-82, 1998.

CLARK, R. M.; SCHWEIKERT, G.; TOOMAJIAN, C.; OSSOWSKI, S.; ZELLER, G.; SHINN, P.; WARTHMANN, N.; HU, T. T.; FU, G.; HINDS, D. A.; CHEN, H.; FRAZER, K. A.; HUSON, D. H.; SCHÖLKOPF, B.; NORDBORG, M.; RÄTSCH, G.; ECKER, J. R.; WEIGE, D. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. **Science**, Washington, v. 317, n. 5836, p. 338-342, 2007.

DAVIS, G. L.; McMULLEN, M. D.; BAYSDORFER, C.; MUSKET, T.; GRANT, D.; K. HOUCHINS. A maize map standard with sequenced core markers, grass genome reference points and 932 expressed sequence tagged sites (ESTs) in a 1736-locus map. **Genetics**, Bethesda, v. 152, n. 3, p. 1137-1172, 1999.

De MASSY, B. Distribution of meiotic recombination sites. **Trends in genetics**, Cambridge, v. 19, n. 9, p. 514-522, 2003.

DOONER. H. K. Genetic fine structure of the *bronze* locus in Maize. **Genetics,** Bethesda, v. 113: p. 1021-1036, 1986.

DROUAUD, J.; CAMILLERI, C.; BOURGUIGNON, P.; CANAGUIER, A.; BÉRARD, A.; VEZON, D.; GIANCOLA, S.; BRUNEL, D.; COLOT, V.; PRUM, B.; QUESNEVILLE, H.; MÉZARD, C. Variation in crossing-over rates along chromosome 4 of Arabidopsis thaliana reveals the presence of meiotic recombination “hot sposts”. **Genome Research**, Cold Spring Harbor, v. 16, p. 106-114, 2006.

DVORAK, J.; YANG, Z.; YOU, F. M.; LUO, M. Deletion polymorphism in wheat chromosome regions with contrasting recombination rates. **Genetics**, Bethesda, v. 168, p. 1665-1675, 2004.

DVORAK, J.; AKHUNOV, E. D. Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploidy evolution in the aegilops-triticum alliance. **Genetics**, Bethesda, v. 171, p. 323-332, 2005.

EWENS, W. J. The sampling theory of selectively neutral alleles. **Theoretical Population Biology**, Maryland, v. 3, n. 1, p. 87-112, 1972.

FEARNHEAD, P.; DONNELLY, P. Estimating recombination rates from population genetic data. **Genetics**, Bethesda, v. 159, p. 1299-1318, 2001.

FEARNHEAD, P.; DONNELLY, P. Approximate likelihood methods for estimating local recombination rates. **Journal of the Royal Statistical Society, Series B (Statistical Methodology)**, Hoboken, v. 64, n. 4, p. 657-680, 2002.

FEARNHEAD, P.; SMITH, N. G. C. A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. **American journal of human genetics**, Cambridge, v. 77, n. 5, p.781-794, 2005.

FEARNHEAD, P. SequenceLDhot: detecting recombination hotspots. **Bioinformatics**, Oxford, v. 22, n. 24, p. 3061-3066, 2006.

FELSENSTEIN, J.; KUHNER, M. K.; YAMATO, J.; BEERLI, P. IMS Lecture Notes-Monograph Series, 33. In: SEILLIER-MOISEIWITSCH, F. (ed.). **Statistics in Molecular Biology and Genetics**. Hayward: Institute of Mathematical Statistics and American Mathematical Society, 1999, p. 163-185.

FRANSZ, P. F.; ARMSTRONG, S.; de JONG, J. H.; PARNELL, L. D.; van DRUNEY, C.; DEAN, C.; ZABEL, P.; BISSELING, T.; JONES, G. H. Integrated cytogenetic map of chromosome arm 4S of *Arabidopsis thaliana*: structural organization of heterocromatic knob and centromere region. **Cell**, Cambridge, v. 100, n. 3, p. 367-376, 2000.

FU, H.; ZHENG, Z.; DOONER, H. K. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. **Proceedings of The National Academy of Science of USA** , Washington, v. 99, p. 1082-1087, 2002.

GAUT, B. S.; WRIGHT, S. I.; RIZZON, C.; DVORAK, J.; ANDERSON, L. K. Recombination: an underappreciated factor in the evolution of plant genomes. **Nature Reviews Genetics**, New York, v. 8, p. 77-84, 2007.

GREENBAUM, J. A.; PANG, B.; TULLIUS, T. D. Construction of a genome-scale structural map at single-nucleotide resolution. **Genome Research**, Cold Spring Harbor, v. 17, p. 947-953, 2007.

GRIFFITHS, R. C.; MARJORAM, P. An ancestral recombination graph. In: DONNELLY, P.; TAVARÉ, S. (Ed.). **IMA Volume in mathematical population genetics**. New York: Springer-Verlag, p. 257-270, 1996.

GRIFFITHS, R. C.; TAVARÉ, S. Simulating probability distributions in the coalescent. **Theoretical Population Biology**, Maryland, v. 46, n. 2, p. 131-159, 1994a.

GRIFFITHS, R. C.; TAVARÉ, S. Ancestral inference in population genetics. **Statistical science**, Beachwood, v. 9, n. 3, p. 307-319, 1994b.

GRIFFITHS, R. C.; TAVARÉ, S. Sampling theory for neutral alleles in a varying environment. **Philosophical transactions of royal society of London B**, London, v. 344, n. 1310, p. 403-410, 1994c.

HAYASHI, T.; IWATA, H. EM algorithm for Bayesian estimation of genomic breeding values. **BMC Genetics**, London, v. 11, n. 3, 2010.

HELLENTHAL, G.; STEPHENS, M. Insights into recombination from population genetic variation. **Current opinion in genetics & development**, Hoboken, v. 16, n. 6, p. 565-572, 2006.

HELLENTHAL, G.; STEPHENS, M. msHOT: modifying Hudson's ms simulator to incorporate croosover and gene conversion hotspots. **Bioinformatics**, Oxford, v. 23, p. 520-521, 2007.

HILL, W. G.; ROBERTSON, A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, Heidelberg, v. 38, p. 226-231, 1968.

HUDSON, R. R. Properties of a neutral allele model with intragenic recombination. **Theoretical Population Biology**, Maryland, v. 23, n. 2, p. 183-201, 1983.

HUDSON, R. R. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. **Genetics**, Bethesda, v. 109, p. 611-631, 1985.
HUDSON, R. R.; KAPLAN, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequence. **Genetics**, Bethesda, v. 111, p. 147-164, 1985.

HUDSON, R. R. Two-locus sampling distributions and their application. **Genetics**, Bethesda, v. 159, p. 1805-1817, 2001.

IRGSP (2005). International Rice Genome Sequencing Project. The map-based sequence of the rice genome. **Nature**, New York, v. 436, p. 793-800, 2005.

JELESKO, J. G.; CARTERA, K.; THOMPSONA, W.; KINOSHITA, Y.; GRUISSEM, W. Meiotic recombination between paralogous RBCSB genes on sister chromatids of *Arabidopsis thaliana*. **Genetics**, Bethesda, v. 166, p. 947-957, 2004.

JONES, L. E.; RYBKA, K.; LUKASZEWSKI, A. J. The effect of a deficiency and a deletion on recombination in chromosome 1BL in wheat. **Theoretical and Applied Genetics**, Heidelberg, v. 104, n. 6-7, p. 1204-1208, 2002.

KENDAL, W.; SUOMELA, B. Large-scale genomic correlations in Arabidopsis thaliana relate to chromosomal structure. **BMC Genomics**, London, v. 6, n. 82, p. 1-7, 2005.

KIM, J. S.; ISLAM-FARIDI, M. N.; KLEIN, P. E.; STELLY, D. M.; PRICE, H. J.; KLEIN, R. R.; MULLET, J. E. Comprehensive molecular cytogenetic analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. **Genetics**, Bethesda, v. 171, n. 4, p. 1963-1976, 2005.

KIM, S.; PLAGNOL, V.; HU, T. T.; TOOMAJIAN, C.; CLARK, R. M.; OSSOWSKI, S.; ECKER, J. R.; WEIGEL, D.; NORDBORG, M. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. **Nature genetics**, New York, v. 39, n. 9, p. 1151-1155, 2007.

KINGMAN, J. F. C. The coalescent. **Stochastic Processes and Their Applications**, Maryland, v. 13, n. 3, p. 235-248, 1982.

KIRKPATRICK, D. T.; WANG, Y.; DOMINSKA, M.; GRIFFITH, J. D.; PETES, T. D. Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes. **Molecular and Celular Biology**, Washington, v. 19, n. 11, p. 7661-7671, 1999.
KUHNER, M. K.; YAMATO, J.; FELSENSTEIN, J. Maximum likelihood estimation of recombination rates from population data. **Genetics**, Bethesda, v. 156, p. 393-1401, 2000.

LEFEBVRE, J. F.; LABUDA, D. Fraction of informative recombinations: a heuristic

approach to analyze recombination rates. **Genetics**, Bethesda, v. 178, p. 2069-2079, 2008.

LEWONTIN, R. C.; KOJIMA, K. The evolutionary dynamics of complex polymorphisms. **Evolution**, Lawrence, v. 14, n. 4, p. 458-472, 1960.

LI, N.; STEPHENS, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. **Genetics**, Bethesda, v. 165, p. 2213-2233, 2003.

LI, Y.; BOREVITZ, J. O. ***Arabidopsis thaliana*** **hap map**. Chigago: University of Chigago/ Department of Ecology and Evolution, 2009. Disponível em <http://borevitzlab.uchicago.edu/Members/yanli1/core360-1>. Acesso em: 15 jul. 2009.

LI, W; HOLSTE, D. An unusual 500,000 bases long oscillation of guanine and cytosine content in human chromosome 21. **Computational Biology and Chemistry**, Oxford, v. 28, n. 5-6, p. 393-399, 2004.

LICHTEN, M.; GOLDMAN, A. S. H. Meiotic recombination hotspots. **Annual Review in Genetics**, Bethesda, v. 29, p. 423-444, 1995.

LIN, Z.; KONG, H.; NEI, M.; MA, H. Origins and evolution of the recA/RAD51 gene family: Evidence for ancient gene duplication and endosymbiotic gene transfer. **Proceedings of the National Academy of Sciences**, Washington, v. 103, n. 27, p. 10328-10333, 2006.

LIÓ, P. Wavelets in bioinformatics and computational biology: state of art and perspactives. **Bioinformatics**, London, v. 19, n. 1, p. 2-9, 2003.

LOCKTON, S.; GAUT, B. S. Plant conserved non-coding sequences and paralogue evolution. **Trends in Genetics**, Cambridge, v. 21, p. 60-65, 2005.

LYSAK, M. A.; BERR, A.; PECINKA, A.; SCHMIDT, R.; McBREEN, L.; SCHUBERT, I. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. **Proceedings National Academy of Science of USA**, Washington, v. 103, p. 5224-5229, 2006.

MARAIS, G.; CHARLESWORTH, B.; WRIGHT, S. I. Recombination and base compositon: the case of highly self-fertilizing plant *Arabidopsis thaliana*. **Genome Biology**, London, v. 5, n. 7, p. 451-459, 2004.

MATSUO, M.; ITO, Y.; YAMAUCHI, R.; OBOKATA, J. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. **Plant Cell**, Rockville, v. 17, p. 665-675, 2005.
MCVEAN, G. A. T.; AWADALLA, P.; FEARNHEAD, P. A coalescent-based method for detecting and estimating recombination from gene sequences. **Genetics**, Bethesda, v. 160, p. 1231-1241, 2002.

MCVEAN, G. A. T.; MYERS, S. R.; HUNT, S.; DELOUKAS, P.; BENTLEY, D. R.; DONNELLY, P. The fine-scale structure of recombination rate variation in the human

genome. **Science**, Washington, v. 304, n. 5670, p. 581-584, 2004.

MERCIER, R. ; JOLIVET, S. ; VEZON, D. ; HUPPE, E. ; CHELYSHEVA, L.; GIOVANNI, M.; NOGUÉ, F.; DOUTRIAUX, M. ; HORLOW, C.; GRELON, M. Two meiotic crossover classes cohabit in Arabidopsis: one is dependent on MER3, whereas the other one is not. **Current Biology**, Maryland, v. 15, n. 8, p. 692-701, 2005.
MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Bethesda, v. 157, n. 4, p. 1819-1829, 2001.

MÉZARD. C. Meiotic recombination hotspots in plants. **Biochemical Society Transactions**, London, v. 34, n. 4, p. 531-534, 2006.

MÉZARD, C.; VIGNARD. J.; DROUAD, J.; MERCIER, R. The road to crossovers: plants have their say. **Trends in genetics**, Cambridge, v. 3, n. 2, p. 93-99, 2007.

MYERS, S. R.; GRIFFITHS, R. C. Bounds on the minimum number of recombination events in a sample history. **Genetics**, Bethesda, v. 163, p. 375-394, 2003.

MYERS, S.; BOTTOLO, L.; FREEMAN, C.; McVEAN, G.; DONNELLY, P. A fine-scale map of recombination rates and hotspots across the human genome. **Science**, Washington, v. 310, p. 321-324, 2005.

MYERS, S.; SPENCER, C. C. A.; AUTON, A.; BOTTOLO, L.; FREEMAN, C.; DONNELLY, P; McVEAN, G. The distribution and causes of meiotic recombination in the human genome. **Biochemical Society Transactions**, London, v. 34, n. 4, p. 526-530, 2006.

NABIROCHKIN, S; OSSOKINA, M; HEIDMAN, T. A nuclear Matrix/Scaffold Attachment Region co-localizes with the Gypsy retrotransposon insulator sequence. **The Journal of Biological Chemistry**, Bethesda, v. 273, n. 4, p. 2473-2479, 1998.

NIELSEN, R. Estimation of population parameters and recombination rates from single polymorphism nucleotides. **Genetics**, Bethesda, v. 154, p. 931-942, 2000.

NORDBORG, M.; HU, T. T.; ISHINO, Y.; JHAVERI, J.; TOOMAJIAN, C.; ZHENG, H.; BAKKER, E.; CALABRESE, P.; GLADSTONE, J.; GOYAL, R.; JAKOBSSON, M.; KIM, S.; MOROZOV, Y.; PADHUKASAHASRAM, B.; PLAGNOL, V.; ROSENBERG, N. A.; SHAH, C.; WALL, J. D.; WANG, J.; ZHAO, K.; KALBFLEISCH, T.; SCHULZ, V.; KREITMAN, M.; BERGELSON, J. The pattern of polymorphism in *Arabidopsis thaliana*. **PLoS Biology**, San Francisco, v. 3, n. 7, p. 1289-1299, 2005.
NORDBORG, M. ***Arabidopsis thaliana* hap map**. Los Angeles: University of Southern California/ Department of Molecular and Computational Biology, 2009. Disponível em <http://walnut.usc.edu/2010/data/SNPs-250k-336.zip>. Acesso em: 24 out. 2009.

PAIGEN, K.; SZATKIEWICZ, J. P.; SAWYER, K.; LEAHY, N.; PARVANOV, E.D.; SIEMON, H. S. N.; GRABER, J. H.; BROMAN,, K. W.; PETKOV, P. M. The recombinational anatomy of a mouse chromosome. **PLoS Genetics**, San Francisco, v. 4, n. 7, p. 1-15, 2008.

PETES, T. D. Meiotic recombination hot spots and cold spots. **Nature Reviews Genetics**, New York, v. 2, p. 360-369, 2001.

QI, L. L.; FRIEBE, B.; GILL, B. S. A strategy for enhancing recombination in proximal regions of chromosomes. **Chromosome Research**, Dordrecht , v. 10, n. 8, p. 645-654, 2002.

RIZZON, C.; PONGER, L.; GAUT, B. S. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. **PLoS Computational Biology**, San Francisco, v. 2, n. 9, p. 989-1000, 2006.

SHRADER, T. E.; CROTHERS, D. M. Artificial nucleosomo positioning sequences. **Proceedings of the National Academy of Sciences of the USA**, Washington, v. 86, n. 19, p. 7418-7422, 1989.

SMITH, G. R.; BODDY, M. N.; SHANAHAN, P.; RUSSELL, P. Fission yeast Mus81·Eme1 holliday junction resolvase is required for meiotic crossing over but not for gene conversion. **Genetics**, Bethesda, v. 165, p. 2289-2293, 2003.

STACK, S. M.; ANDERSON, L. K. Two-dimensional spreads of synaptonemal complexes from solanaceous plants: synapsis in *Lycopersicon esculentum* (Tomato). **American Journal of Botany**, Saint Louis, v. 73, n. 2, p. 264-281, 1986.

STEINER, W. W.; SMITH, G. R. Optimizing the nucleotide sequence of a meiotic recombination hotspot in *Schizosaccharomyces pombe*. **Genetics**, Bethesda, v. 169, p. 1973-1983, 2005a.

STEINER, W. W.; SMITH, G. R. Natural meiotic recombination hot spots in *Schizosaccharomyces pombe* genome successfully predicted from the simple sequence motif M26. **Molecular and Cellular Biology**, Washington, v. 25, n. 20, p. 9054-9062, 2005b.

STEPHENS, J. C. On the frequency of indetectable recombination events. **Genetics**, Bethesda, v. 112, n. 4, p. 923-926, 1986.

STEPHENS, M.; DONNELLY, P. Inference in molecular population genetics. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, Hoboken, v. 62, n. 4, p. 605-655, 2000.

STEPHEN, M.; SMITH, N. J.; DONNELLY, P. A new statistical method for haplotype reconstruction from population data. **American Journal of Human Genetics**, Boston, v. 68, n. 4, p. 978-989, 2001.

STEPHENS, M. Inference under coalescent. In: BALDING, D. J.; BISHOP, M.; CANNINGS, C. (Ed.). **Handbook of statistical genetics**. 3. ed. Chichester: Jonh Wiley & Sons, v. 2, cap. 26, p. 878-908, 2007.

STUMPF, M. P. H.; MCVEAN, G. A. T. Estimating recombination rates from population-genetic data. **Nature reviews genetics**, New York, v. 4, n. 12 , p. 959-968, 2003.

SZOSTAK, J. W.; ORR-WEAVER, T. L.; ROTHSTEIN, R. J.; STAHL, F. W. The double-strand-break repair model for recombination, **Cell**, Cambridge, v. 33, n. 1, p. 25-35, 1983.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. **Genetics**, Bethesda, v. 123, n. 3, p. 585-595, 1989.

TANKSLEY, S. D.; GANAL, M. W.; PRINCE, J. P.; DE-VICENTE, M. C.; BONIERBALE, M. W.; BROUN, P.; FULTON, T. M.; GIOVANNONI, J. J.; GRANDILLO, S.; MARTIN, G. B.; MESSEGUER, R.; MILLER, J. C.; MILLER, L.; PATERSON, A. H.; PINEDA, O.; RODER, M. S.; WING, R. A.; WU, W.; YOUNG, N. D. High density molecular linkage maps of the tomato and potato genomes. **Genetics**, Bethesda, v. 132, p. 1141-1160, 1992.

TENAILLON, M. I.; SAWKINS, M. C.; ANDERSON, L. K.; STACK, S. M.; DOEBLEY, J.; GAUT, B. S. Patterns of Diversity and Recombination Along Chromosome 1 of maize (Zea mays ssp. mays L.). **Genetics**, Bethesda, v. 162, p. 1401-1413, 2002.

THOMAS, J. H. Analysis of homologous gene clusters in Caenorhabditis elegans reveals striking regional cluster domains. **Genetics**, Bethesda, v. 172, p. 127-143, 2006.

WAKELEY, J. **Coalescent theory: an introduction**. Greenwood Village: Roberts & Company, 2006. 220 p.

WALL, J. D. A comparison of estimators of the population recombination rate. **Molecular biology and evolution**, Oxford, v.17, n. 1, p. 156-163, 2000.

WANG, C. R.; HARPER, L.; CANDE, W. Z. High-resolution single-copy gene fluorescence in situ hybridization and its use in the construction of a cytogenetic map of maize chromosome 9. **The Plant Cell**, Rockville, v. 18, p. 529-544, 2006.

WANG, Y.; RANNALA, B. Bayesian inference of fine-scale recombination rates using population genomic data. **Philosophical trasactions of the royal society B (Biological Sciences)**, London, v. 363, n. 1512, p. 3921-3930, 2008.

WANG, Y.; TANG, X.; CHENG, Z.; MUELLER, L.; GIOVANNONI, J.; TANKSLEY, S. D. Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. **Genetics**, Bethesda, v. 172, p. 22529-22540, 2006.

WANG, Y. H.; GELLIBOLIAN, R.; SHIMIZU, M.; WELLS, R. D.; GRIFFITH, J. D. Long CCG triplet repeats exclude nucleosomes: a possible mechanism for the nature of fragile sites in chromosomes. **Journal of Molecular Biology**, Maryland, v. 263, p. 511-516, 1996.

WARBURTON, P. E. ; COOKE, C. A.; BOURASSA, S.; VAFA, O.; SULLIVAN, B. A.; STETTEN, G.; GIMELLI, G.; WARBURTON, D.; TYLER-SMITH, C.; SULLIVAN, K. F.; POIRIER. G. G.; EARNSHAW, W. C. Immunolocalization of CENP-A suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres. **Current Biology**, Maryland Heights, v. 7, n. 11. p. 901-904, 1997.

WEIR, B. S. Inferences about linkage disequilibrium. **Biometrics**, New York, v. 35, n. 1, p. 235-254, 1979.

WHITBY, M. C. Making crossovers during meiosis. **Biochemical Society Transactions**, London, v. 33, n. 6, p. 1451-1455, 2005.

WILLIAMS, B. C.; MURPHY, T. D.; GOLDBERG, M. L.; KARPEN, G. H. Neocentromere activity of structurally acentric mini-chromosomes in *Drosophila*. **Nature Genetics**, New York, v. 18, p. 30-38, 1998.

WIUF, C. On the minimum number of topologies explaining a sample of DNA sequences. **Theoretical population biology**, Oxford, v. 62, n. 4, p. 357-363, 2002.

WRIGHT, S. I.; AGRAWAL, N.; BUREAU, T. E. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. **Genome Research**, Cold Spring Harbor, v.13, p. 1897-1903, 2003.

YAN, H.; JIN, W.; NAGAKI, K.; TIAN, S.; OUYANG, S.; BUELL, C. R.; TALBERT, P. B.; HENIKOFF, S.; JIANG, J. Transcription and histone modifications in the recombination-free region spanning a rice centromere. **The Pant Cell**, Rockville, v. 17, p. 3227-3238, 2005.

YAO, H.; ZHOU, Q.; LI, J.; SMITH, H. ; YANDEAU, M.; NIKOLAU, B. J.; SCHNABLE, P. S. Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. **Proceedings of the National Academy of Science of the USA**, Washington, v. 99, n. 9, p. 6157-6162, 2002.

ZHANG, L.; GAUT, B. S. Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? **Genome Research**, Cold Spring Harbor, v. 13, p. 2533-2540, 2003.

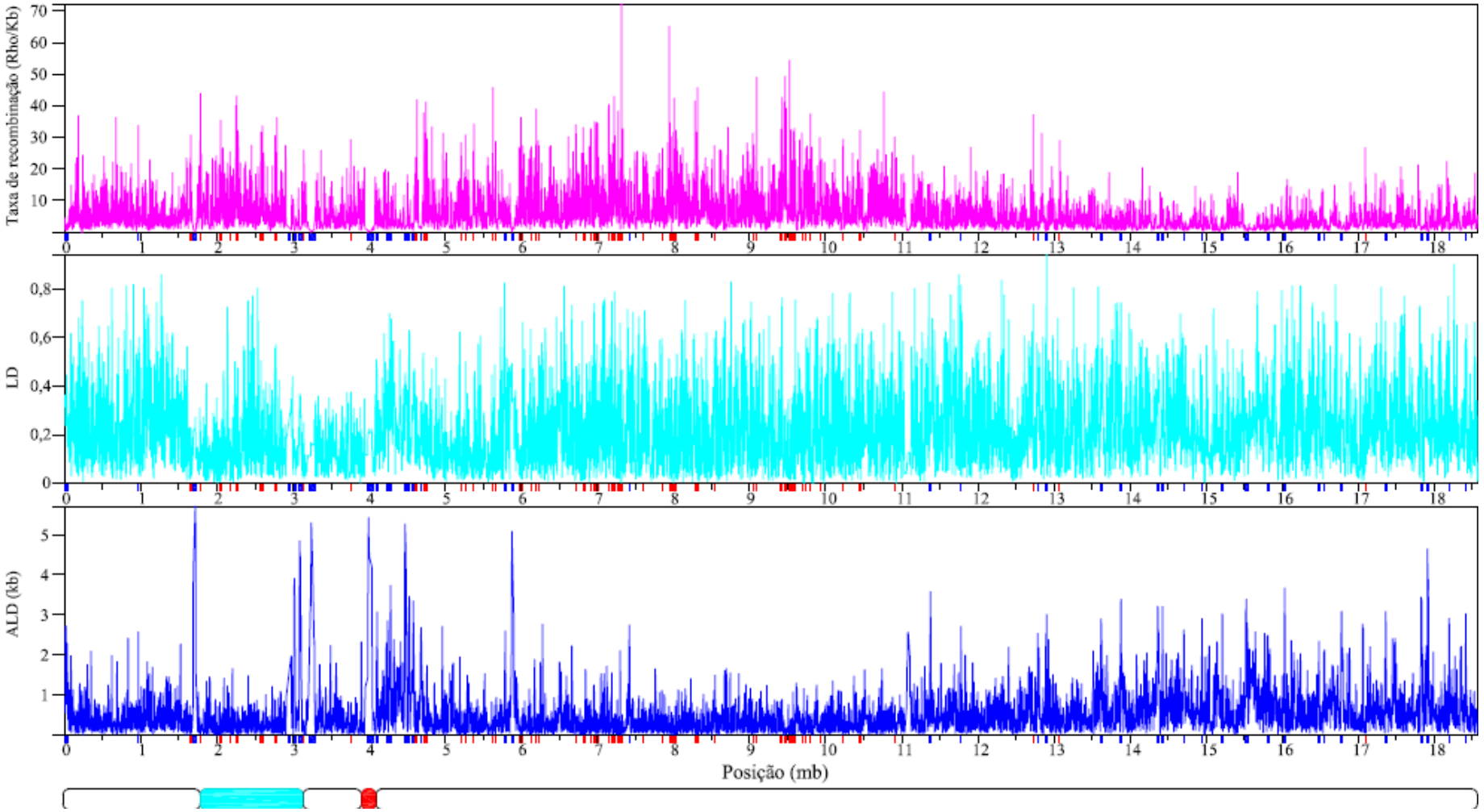
ZIETKIEWICZ, E. ; YOTOVA, V. ; GEHL, D. ; WAMBACH, T.; ARRIETA, I.; BATZER, M.; COLE, D. E. C.; HECHTMAN, P.; KAPLAN, F.; MODIANO, D.; MOISAN, J. P. ; MICHALSKI, R. ; LABUDA, D. Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. **The American Journal of Human Genetics**, Boston, v. 73, p. 994-1015, 2003.

ZIOLKOWSKI, P. A.; BLANC, G.; SADOWSKI, J. Structural divergence of chromosomal segments that arose from successive duplication events in the *Arabidopsis* genome. **Nucleic Acids Research**, Oxford, v. 31, n. 4, p. 1339-1350, 2003.

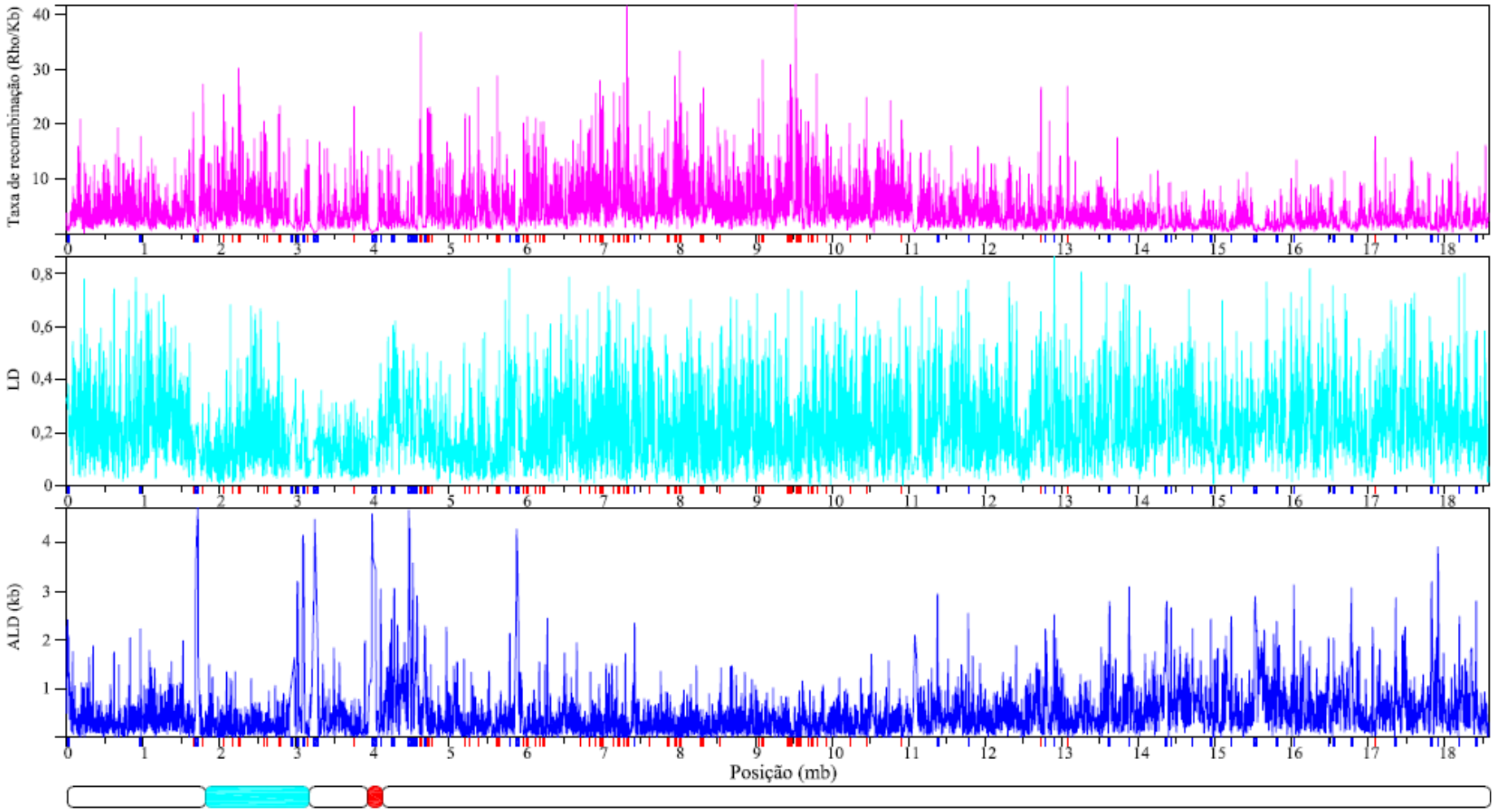
APÊNDICES

Apêndice A

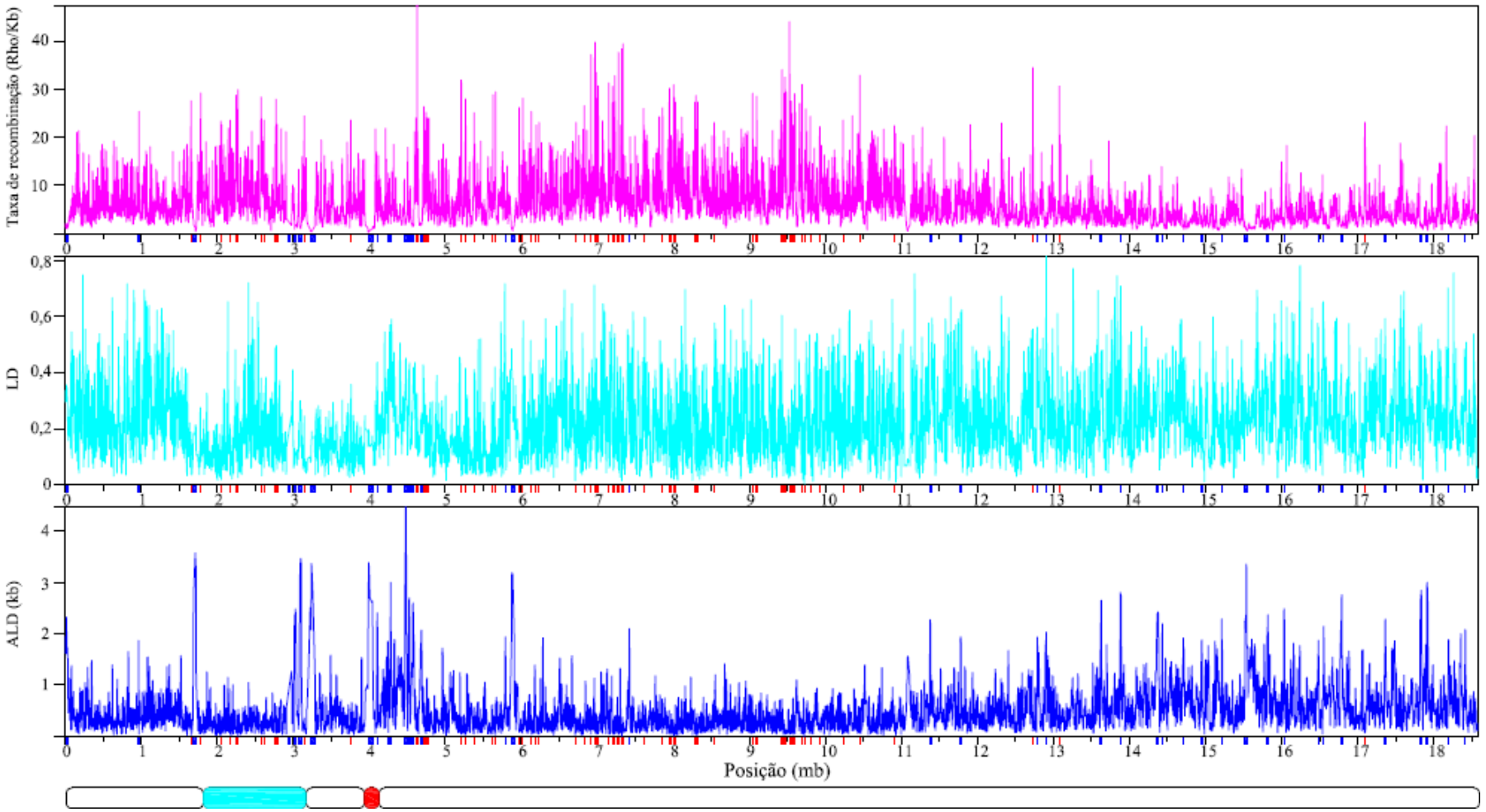
Figuras contendo gráficos da variação da taxa de recombinação populacional ($\boldsymbol{\rho}$), do desequilíbrio de ligação (LD) e do alcance do desequilíbrio de ligação (ALD) ao longo do cromossomo 4 de *Arabidopsis thaliana*, usando-se janelas de tamanho 10 (Apêndice A-1), 12 (Apêndice A-2) e 16 (Apêndice A-3) locos.



Apêndice A-1. Variação da taxa de recombinação populacional (***ρ***), do e do desequilíbrio de ligação (LD) e do alcance do desequilíbrio de ligação (ALD) ao longo do cromossomo 4 de *A.rabidopsis thaliana.* Na parte superior, em magenta, a variação de ***ρ*** (número de *crossover* por kb). Ao centro, em ciano, a variação de LD e, na parte inferior, em azul, a variação de ALD (kb). As estimativas de Rho, LD e do ALD foram calculadas usando uma janela igual a 10 locos. Junto à parte inferior da linha da escala do eixo x encontram-se pequenos retângulos que representam os fragmentos com valores tão ou mais extremos que os valores de Corte(99%) arbitrados: os azuis representam fragmentos que apresentam valores extremos de ALD; os vermelhos representam os fragmentos que apresentaram valores extremos de ***ρ***. Na base da figura encontra-se uma representação esquemática do cromossomo 4 de *A. thaliana*, na mesma escala do eixo x. A parte vermelha representa a região centromérica, e, a ciano oum *knob* heterocromático.



Apêndice A-2. Variação da taxa de recombinação populacional (***ρ***), do e do desequilíbrio de ligação (LD) e do alcance do desequilíbrio de ligação (ALD) ao longo do cromossomo 4 de *A.rabidopsis thaliana.* Na parte superior, em magenta, a variação de ***ρ*** (número de *crossover* por kb). Ao centro, em ciano, a variação de LD e, na parte inferior, em azul, a variação de ALD (kb). As estimativas de Rho, LD e do ALD foram calculadas usando uma janela igual a 12 locos. Junto à parte inferior da linha da escala do eixo x encontram-se pequenos retângulos que representam os fragmentos com valores tão ou mais extremos que os valores de Corte(99%) arbitrados: os azuis representam fragmentos que apresentam valores extremos de ALD; os vermelhos representam os fragmentos que apresentaram valores extremos de ***ρ***. Na base da figura encontra-se uma representação esquemática do cromossomo 4 de *A. thaliana*, na mesma escala do eixo x. A parte vermelha representa a região centromérica, e, a ciano oum *knob* heterocromático.



Apêndice A-3. Variação da taxa de recombinação populacional (***ρ***), do e do desequilíbrio de ligação (LD) e do alcance do desequilíbrio de ligação (ALD) ao longo do cromossomo 4 de *A.rabidopsis thaliana.* Na parte superior, em magenta, a variação de ***ρ*** (número de *crossover* por kb). Ao centro, em ciano, a variação de LD e, na parte inferior, em azul, a variação de ALD (kb). As estimativas de Rho, LD e do ALD foram calculadas usando uma janela igual a 16 locos. Junto à parte inferior da linha da escala do eixo x encontram-se pequenos retângulos que representam os fragmentos com valores tão ou mais extremos que os valores de Corte(99%) arbitrados: os azuis representam fragmentos que apresentam valores extremos de ALD; os vermelhos representam os fragmentos que apresentaram valores extremos de ***ρ***. Na base da figura encontra-se uma representação esquemática do cromossomo 4 de *A. thaliana*, na mesma escala do eixo x. A parte vermelha representa a região centromérica, e, a ciano oum *knob* heterocromático.