DOI: 10.1534/genetics.105.042978

Modeling Haplotype Block Variation Using Markov Chains

G. Greenspan[1] and D. Geiger

Computer Science Department, Technion, Haifa 32000, Israel

Manuscript received March 7, 2005
Accepted for publication November 23, 2005

ABSTRACT

Models of background variation in genomic regions form the basis of linkage disequilibrium mapping methods. In this work we analyze a background model that groups SNPs into haplotype blocks and represents the dependencies between blocks by a Markov chain. We develop an error measure to compare the performance of this model against the common model that assumes that blocks are independent. By examining data from the International Haplotype Mapping project, we show how the Markov model over haplotype blocks is most accurate when representing blocks in strong linkage disequilibrium. This contrasts with the independent model, which is rendered less accurate by linkage disequilibrium. We provide a theoretical explanation for this surprising property of the Markov model and relate its behavior to allele diversity.

GENETIC mapping studies based on linkage disequilibrium (LD) require a model of the background variation in the region being examined. The studies look for markers at which the norms of this model are violated by a set of diseased individuals, inferring that any such markers are likely to be close to a disease-causing allele. While many LD mapping methods do not explicitly refer to such a background model, it exists nonetheless as an underlying assumption. For example, a χ^2-correlation test between disease status and the allele frequencies at individual SNPs assumes that the alleles at each SNP are distributed independently.

Recent work on human genomic variation suggests that it is fruitful to group SNP markers together into haplotype blocks (Daly *et al.* 2001; Goldstein 2001; Patil *et al.* 2001; Gabriel *et al.* 2002; Wall and Pritchard 2003). These blocks are thought to be delineated by recombination hotspots, small areas in which the probability of recombination is far higher than that in the surrounding regions (Jeffreys *et al.* 2000, 2001; Templeton *et al.* 2000; Wang *et al.* 2002; Arnheim *et al.* 2003; Phillips *et al.* 2003; Twells *et al.* 2003). The low probability of recombination inside each block means that the alleles at the SNPs within are passed together from one generation to the next. Therefore, each haplotype block can be considered as a single marker, with the set of alleles at the SNPs in the block constituting its allele (Cardon and Abecasis 2003; Tishkoff and Verrelli 2003). It is hoped that haplotype blocks will enable fewer SNP markers to be genotyped during LD mapping studies, since a small number of haplotype-tagging SNPs (htSNPs) can be used to identify the common alleles of each block (Johnson *et al.* 2001; Zhang *et al.* 2002; Sebastiani *et al.* 2003). The International Haplotype Mapping Project (HapMap) is currently producing a high-density haplotype map of the human genome for several target populations, to enable the efficient selection of htSNPs (International HapMap Consortium 2003).

It is not practical or desirable to represent the background variation over SNP or block markers in a large chromosomal region using a full joint distribution. A model must also infer something about the structure of the distribution, so that it is sufficiently robust to deal with additional individuals or a future generation. Since models of background variation are generally inferred from a small sample of haplotypes, many haplotypes present in the population will not appear in the sample obtained.

The most obvious approximation of the full joint distribution is a model that assumes that all markers are independent, *i.e.*, that the probability of a haplotype is the product of the frequencies of each allele within. This type of model is common and constitutes an implicit assumption in many LD mapping studies. The independent model has the advantage of requiring a small number of parameters, namely the frequencies for each allele. However, this model breaks down when representing the variation over short distances, since markers that are close together tend to exhibit a high degree of linkage disequilibrium that cannot be captured by an independent approximation.

In this article, we focus on a different model, namely the Markov chain. In the Markov model, the probability of each allele at a marker is conditional on the allele present at the previous marker. This model is able to represent some of the correlations that exist in a genomic region, while still keeping to a linear number of

[1] *Corresponding author:* Computer Science Department, Technion, Technion City, Haifa 32000, Israel. E-mail: gdg@cs.technion.ac.il

parameters. For example, when modeling block markers with four possible alleles, the Markov model will require four times as many parameters as the independent model. Many of the existing methods for partitioning regions into haplotype blocks include a Markov chain in their models (Daly *et al.* 2001; Anderson and Novembre 2003; Greenspan and Geiger 2004a; Kimmel and Shamir 2004). Several published approaches to LD mapping also use a Markov chain to represent the background variation over blocks or individual SNPs (McPeek and Strahs 1999; Morris *et al.* 2000; Liu *et al.* 2001; Morris *et al.* 2002; Greenspan and Geiger 2004b).

For any given joint distribution, a Markov approximation will clearly be more accurate than an independent approximation, since it has more parameters available for optimization. However, we describe in this article an important additional property of the Markov approximation that we consider surprising. When used to model haplotype blocks, the Markov approximation is most accurate in the presence of high levels of linkage disequilibrium. Consequently, the Markov model is more accurate for blocks that are close together than for those that are far apart. We also show that when modeling individual SNPs instead of haplotype blocks, this property of the Markov model is not exhibited. In other words, a Markov model over haplotype blocks provides a uniquely accurate way to represent background genomic distributions at high resolution. This result justifies previous work that uses a Markov model over haplotype blocks for both haplotype resolution and LD mapping (Greenspan and Geiger 2004a,b).

MATERIALS AND METHODS

Independent and Markov approximations: Consider a genomic region that contains l markers, placed at physical locations $z_1, \ldots, z_l$ along the chromosome (measured in base pairs). Each marker $j = 1, \ldots, l$ has r_j alleles, labeled $1, \ldots, r_j$. We consider a population in Hardy–Weinberg equilibrium, so the background variation for the region is given in terms of a joint distribution over haplotype frequencies (Hardy 1908). Let $P(x_1, \ldots, x_l)$ be the frequency of haplotype $x_1, \ldots, x_l$ in the population, where each x_j takes the values $1, \ldots, r_j$.

Under the independent model, each marker is assumed to be independent. The maximum-likelihood independent approximation $T(x_1, \ldots, x_l)$ of the joint distribution P is

$$T(x_1, \ldots, x_l) = \prod_{i=1}^{l} P(x_i),$$

where

$$P(x_i) = \sum_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_l} P(x_1, \ldots, x_l).$$

Under the Markov model, the distribution for each marker is dependent on the allele present at the preceding marker. The maximum-likelihood Markov approximation $Q(x_1, \ldots, x_l)$ of the joint distribution is

$$Q(x_1, \ldots, x_l) = P(x_1) \prod_{i=1}^{l-1} P(x_{i+1} \mid x_i),$$

where

$$P(x_{i+1} \mid x_i) = \frac{\sum_{x_1, \ldots, x_{i-1}, x_{i+2}, \ldots, x_l} P(x_1, \ldots, x_l)}{\sum_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_l} P(x_1, \ldots, x_l)}.$$

Error measures: Given a distance d and a number $n \geq 3$ of markers, we generate statistics $Y_{d,n}$ and $Z_{d,n}$ to quantify the average error of the independent and Markov approximations, respectively, over a chromosome or large genomic region. We set a minimum of $n = 3$ since a Markov model can represent any joint distribution over one or two loci perfectly, rendering our measure meaningless.

The statistics $Y_{d,n}$ and $Z_{d,n}$ for a genomic region are generated by averaging the respective sets of statistics $Y_{d,n}(j)$ and $Z_{d,n}(j)$ over all valid start markers j within that region. Every value of $Y_{d,n}$ and $Z_{d,n}$ in our results was calculated by averaging hundreds or thousands of these individual measurements. Each statistic $Y_{d,n}(j)$ or $Z_{d,n}(j)$ measures the error of the independent or Markov approximation over n markers, where the first marker $j_1 = j$ and the other markers $j_2, \ldots, j_n$ are chosen to be spread approximately evenly over total distance d. Each marker j_i is selected to minimize $|z_{j_i} - z_{j_1} - d \cdot (i-1)/(n-1)|$. If any two of the marker indexes $j_1, \ldots, j_n$ are identical, we conclude that there is insufficient marker density for n, d, and j. In this case, j is not a valid start marker and we omit $Y_{d,n}(j)$ and $Z_{d,n}(j)$ from their respective averages.

We set $Y_{d,n}(j) = \|P(x_{j_1}, \ldots, x_{j_n}) - T(x_{j_1}, \ldots, x_{j_n})\|$, the variation distance between the observed joint distribution P and the independent approximation T for markers $j_1, \ldots, j_n$. Similarly, we set $Z_{d,n}(j) = \|P(x_{j_1}, \ldots, x_{j_n}) - Q(x_{j_1}, \ldots, x_{j_n})\|$, the variation distance between P and the Markov approximation Q. The variation distance between two distributions is defined as follows:

$$\begin{aligned} &\|A(z_1, \ldots, z_n) - B(z_1, \ldots, z_n)\| \\ &\quad = \frac{1}{2} \sum_{z_1, \ldots, z_n} |A(z_1, \ldots, z_n) - B(z_1, \ldots, z_n)|. \end{aligned}$$

This measure is also known as the total variational distance, Kolmogorov distance, or L_1 distance. It has an intuitive definition as the total amount of probability mass that must be moved to make one distribution equal to the other. For example, $\|P - T\|$ is the percentage of the population distributed as P that is misrepresented by the independent approximation T.

The variation distance between the joint distribution P and its independent approximation T is closely related to the D measure of linkage disequilibrium for two biallelic markers. Consider two markers A and B, each with two alleles a_1, a_2, b_1, and b_2 at frequencies p_1, p_2, q_1, and q_2, respectively. Let p_{11}, p_{12}, p_{21}, and p_{22} be the respective frequencies of the four gametes a_1b_1, a_1b_2, a_2b_1, and a_2b_2. The linkage disequilibrium measure D is defined as $D = p_{11} - p_1q_1 = p_1q_2 - p_{12} = p_2q_1 - p_{21} = p_{22} - p_2q_2$ (Devlin and Risch 1995). For example, if A and B are in perfect linkage equilibrium, then $p_{11} = p_1q_1$, $p_{12} = p_1q_2$, $p_{21} = p_2q_1$, and $p_{22} = p_2q_2$, and so $D = 0$. By comparison, the variation distance between P and T is

$$\begin{aligned} &\|P - T\| \\ &\quad = \tfrac{1}{2}(|p_{11} - p_1q_1| + |p_{12} - p_1q_2| + |p_{21} - p_2q_1| + |p_{22} - p_2q_2|) \\ &\quad = \tfrac{1}{2}(|D| + |D| + |D| + |D|) = 2|D|. \end{aligned}$$

Thus, for two biallelic markers, the variation distance between the joint distribution P and its independent approximation T is twice the absolute value of D.

HapMap analysis: We used the October 2004 data release of HapMap to profile the error rates of the independent and Markov approximations for the human genome (International HapMap Consortium 2003). We inferred the transmitted and untransmitted haplotypes from both parents in the 30 CEPH trios, so that 120 haplotypes were examined for each of the 22 autosomes. Haplotype alleles that could not be determined were left as unknown. This occurred at sites for which (a) a genotype was absent, (b) a Mendelian error was detected, or (c) all three members of the trio were heterozygous.

We examined the HapMap data using three different approaches: (a) treating each SNP as an individual marker, (b) grouping the SNPs into haplotype blocks according to various criteria, and (c) grouping fixed numbers of adjacent SNPs into arbitrary blocks.

For the first approach, each SNP marker had $r_j = 2$ alleles, since all SNP markers genotyped in the HapMap are biallelic. Trivially, z_j was set to the physical location of each SNP.

For the second approach, we used two programs, HaploBlock and HaploBlockFinder, to partition the SNP data for each chromosome into l blocks (Zhang and Jin 2003; Greenspan and Geiger 2004b). Each block inferred was considered as a marker and the variants of that block as the marker's alleles. The physical location z_j of each block j was set to the midpoint of the chromosomal section containing the SNPs within.

HaploBlock uses a statistical model-fitting criterion to infer the most suitable block partition for a genomic region (Greenspan and Geiger 2004b). When inferring blocks with HaploBlock, we removed the dependencies between adjacent ancestor variables in the statistical model, to prevent a potential bias in favor of the Markov approximation. We inferred three full HaploBlock models from the HapMap data, with a maximum of four, six, and eight ancestral haplotypes per block, respectively. The HaploBlock statistical model also allows for recent mutations, so some of the haplotypes observed in a block might differ slightly from their inferred ancestors.

HaploBlockFinder offers a number of different criteria for inferring block paritions (Zhang and Jin 2003). We chose the commonly used chromosomal coverage criterion. This criterion defines a block as a region in which a certain percentage of the chromosomes can be covered by four haplotypes, with no additional mutations. We inferred three full HaploBlockFinder partitions from the HapMap data, with percentage thresholds of 70, 80, and 90%, respectively. As this percentage threshold increases, more of the haplotypes within each block must be covered by four common variants, so less variation is permitted overall.

For the third approach, we grouped sets of up to six adjacent SNPs into block markers, without using any additional criterion. The alleles of each marker were defined by the observed combinations of alleles at the SNPs within. The goal of this approach was to determine whether the results observed for haplotype blocks are specific to the criteria used or whether similar results are observed for such groupings of SNPs.

Recall that we omit values $Y_{d,n}(j)$ and $Z_{d,n}(j)$ from the averages $Y_{d,n}$ and $Z_{d,n}$ if n markers are not available with roughly equal spacing over distance d starting at marker j. We also omitted the statistics $Y_{d,n}(j)$ and $Z_{d,n}(j)$ if less than half of the 120 haplotypes could be used for sites $j_1, \ldots, j_n$, due to missing genotypes or haplotype uncertainty. For the SNP analyses, a haplotype could not be used if one of the alleles at sites $j_1, \ldots, j_n$ was not known. For the block-based analyses, a haplotype could not be used if one of the sets of block alleles could not be assigned to a specific block variant with $>$50% certainty.

TABLE 1

Summary for each chromosome of SNPs and the HaploBlock model with up to four variants

		SNPs		Haplotype blocks	
Chromosome	Length	Count	Mean spacing	Count	Mean spacing
1	245,416	47,618	5.15	4,587	53.47
2	243,363	68,659	3.54	5,604	43.42
3	199,162	43,880	4.54	4,132	48.21
4	191,628	42,213	4.54	4,059	47.19
5	180,747	47,651	3.79	4,070	44.41
6	170,674	37,013	4.61	3,567	47.84
7	158,508	35,460	4.47	3,354	47.24
8	146,201	44,926	3.25	3,536	41.34
9	136,199	30,902	4.41	2,812	48.42
10	134,982	29,391	4.59	2,972	45.38
11	134,292	37,281	3.60	3,106	43.24
12	131,969	34,894	3.78	3,093	42.66
13	96,204	24,138	3.99	2,201	43.68
14	87,070	24,104	3.61	2,098	41.51
15	81,870	22,762	3.60	1,966	41.61
16	90,025	21,516	4.18	2,010	44.76
17	81,652	21,511	3.80	1,950	41.89
18	76,096	20,545	3.70	1,866	40.68
19	63,742	15,265	4.18	1,564	40.69
20	63,623	10,794	5.89	1,402	45.36
21	36,965	17,071	2.17	1,279	27.62
22	34,877	15,520	2.25	1,313	26.41
Overall	2,785,266	693,114	4.02	62,541	44.49

All distances are in kilobases.

RESULTS

Summary of models: Table 1 summarizes the SNP loci examined for each chromosome, as well as the characteristics of the HaploBlock statistical models inferred with up to four variants per block. Table 1 shows that the average SNP spacing over all 22 chromosomes is 4.02 kb, whereas the average spacing between the blocks is 44.49 kb. These values provide a rough lower bound on the distances d and n that can be examined usefully for the respective models, since n markers spread over distance d must be spaced at least $d/(n-1)$ apart to be included in the averages $Y_{d,n}$ and $Z_{d,n}$. Table 2 compares the average values over all 22 chromosomes for the six HaploBlock and HaploBlockFinder models. Table 2 also shows the average number of variants inferred per block for each model.

Distance profiles: We assessed how the error rates of the independent and Markov approximations varied over different distances d. The distance profiles were generated by calculating average values of $Y_{d,n}$ and $Z_{d,n}$ over the entire autosome for values of $3 \le n \le 5$.

We first examine the results for the HaploBlock model with up to four variants per block. Figure 1 shows the error measures $Z_{d,n}$ for the Markov approximation for this model over different distances d. Values are

TABLE 2

Summary over all chromosomes of different haplotype block models

Model	Block count	Mean spacing (kb)	Mean variants
HaploBlock max 4 variants	62,541	44.49	3.60
HaploBlock max 6 variants	48,749	57.07	4.84
HaploBlock max 8 variants	43,364	64.16	5.68
HaploBlockFinder 90% coverage	88,038	31.61	5.75
HaploBlockFinder 80% coverage	53,207	52.34	8.79
HaploBlockFinder 70% coverage	39,575	70.31	11.51

shown relative to $Z_{d,n}$ at long distances, where linkage disequilibrium is minimal. These baseline error measures $Z_{d,n}$ are 0.135, 0.310, and 0.515 for $n = 3$, 4, and 5, respectively. To avoid a bias at short distances toward genomic regions with particularly high levels of variation, the graph in Figure 1 shows only the average for distances $d \geq 100$ kb for which at least 75% of the values $Z_{d,n}(j)$ could be generated.

The graph in Figure 1 highlights our core observation—that the Markov approximation performs best for haplotype blocks that are close together and between which there are high levels of linkage disequilibrium. For example, for $n = 4$ blocks spread over $d = 350$ kb, the Markov approximation shows an 8% improvement compared to 4 blocks spread over the longest distance. For $n = 5$ blocks, the improvement is 15%. Figure 1 also shows that the relationship between distance and accuracy is not monotonic—at intermediate distances, the approximation performs worse than at both shorter and longer distances. This phenomenon can be seen most clearly for $n = 3$ blocks, where the average accuracy of the Markov approximation at $d = 350$ kb is equal to that at long distances, but is less accurate at distances in between.

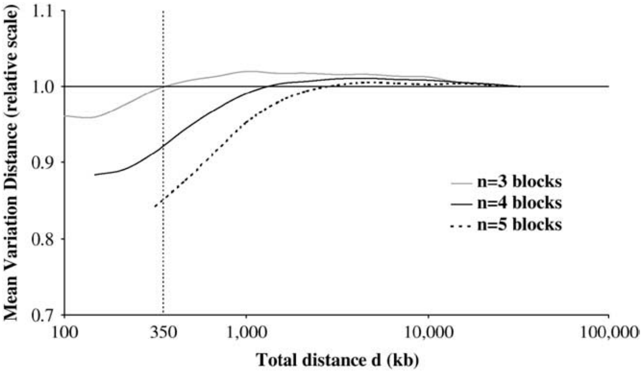

FIGURE 1.—Distance profile of Markov approximation for HaploBlock blocks.

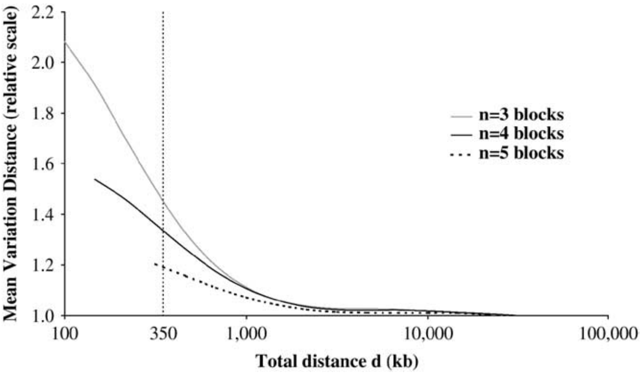

FIGURE 2.—Distance profile of independent approximation for HaploBlock blocks.

Figure 2 shows the corresponding error measures $Y_{d,n}$ for the independent approximation over different distances d. In contrast to Figure 1, this graph shows a monotonic decrease in the independent approximation's error with physical distance. This reflects the fact that the accuracy of the independent approximation improves as the linkage disequilibrium between blocks decreases. One would naturally expect the Markov approximation to behave similarly, yet the results in Figure 1 show otherwise. The values in Figure 2 are shown relative to baseline error measures $Y_{d,n}$ at long distances of 0.194, 0.366, and 0.565 for $n = 3$, 4, and 5, respectively. The baseline increases with the number n of markers due to the increase in the cardinality of distribution $P(x)$, which represents 4^n different haplotypes for blocks with four alleles.

We now compare the results from this HaploBlock model with the approach where each SNP is treated as an individual marker with two alleles. Figure 3 compares the distance profiles of both the Markov and independent approximations for the two approaches, using $n = 4$ in all cases. This graph shows that, when modeling individual SNPs, both the independent and the Markov approximations perform best over longer distances, *i.e.*, where there is less linkage disequilibrium between the

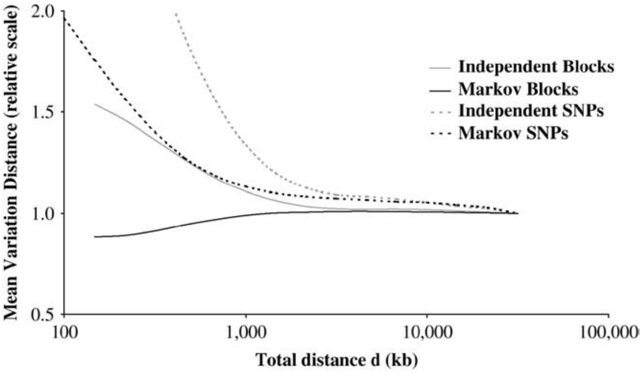

FIGURE 3.—Comparison of distance profiles for HaploBlock blocks and individual SNPs.

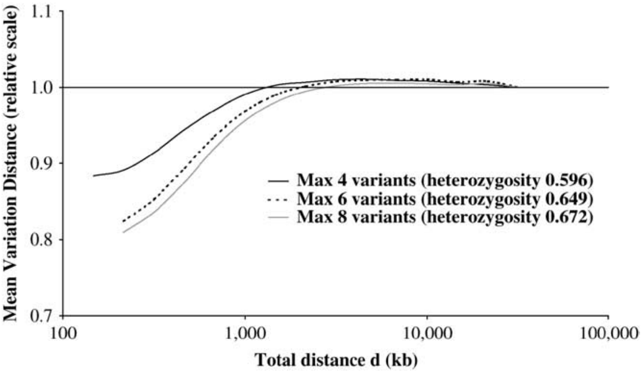

FIGURE 4.—Comparison of Markov distance profiles for HaploBlock models.

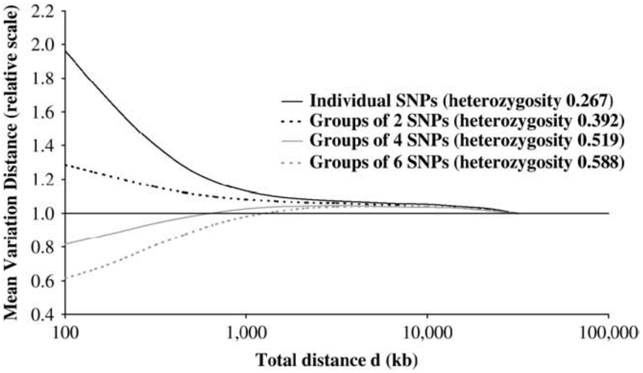

FIGURE 6.—Comparison of Markov distance profiles for SNP groupings.

markers modeled. In other words, the Markov model performs best at short distances only when used with haplotype blocks. As explained later, this difference in behavior between blocks and SNPs is related to the difference in allele diversity.

Figure 4 compares the Markov approximation profiles for the three HaploBlock models with up to four, six, and eight variants per block. Figure 4 shows that, as the number of variants per block is allowed to increase, the improvement in the Markov approximation at short distances becomes more pronounced. In other words, as the allele diversity of the blocks increases, the behavior of the Markov approximation becomes even less like that for individual SNPs. The values in Figure 4 are relative to baseline $Y_{d,n}$ measures of 0.310, 0.442, and 0.506 for four, six, and eight variants, respectively.

Figure 5 compares the Markov approximation profiles for the three HaploBlockFinder partitions. Recall that the threshold specifies the percentage of the variation within each block that can be covered by four common variants. Figure 5 shows that, as the threshold is relaxed to allow more variation within each block, there is more improvement in the Markov approximation at short distances. Once again, this shows the effect of allele diversity. The values in Figure 5 are relative to baseline measures of 0.397, 0.550, and 0.631 for coverage thresholds of 70, 80, and 90%, respectively.

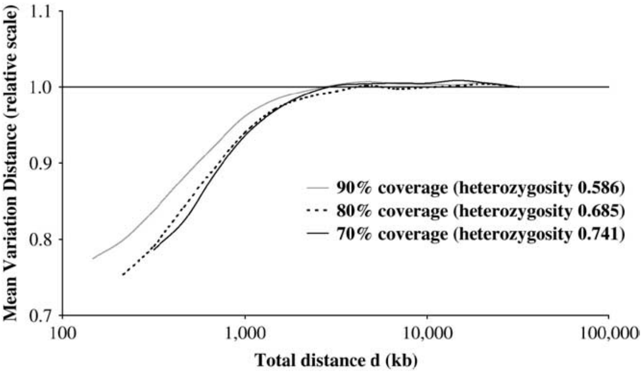

FIGURE 5.—Comparison of Markov distance profiles for HaploBlockFinder models.

Finally, Figure 6 compares the Markov approximation profiles for blocks based on arbitrary groupings of SNPs. Figure 6 shows the effect of increasing the number of SNPs per group on the performance of the Markov approximation. Whereas groups of one or two SNPs perform worse at short distances than at long distances, this relationship is reversed for groups of four SNPs or more. The values in Figure 6 are relative to baselines of 0.051, 0.123, 0.184, and 0.294 for one, two, four, and six SNPs, respectively.

The curves in Figures 4–6 are labeled with the average heterozygosity of their respective inferred markers. Each set of curves shows a clear correlation between increased marker heterozygosity and the increased accuracy of the Markov approximation at short distances. As explained later, this relationship stems from the effects of marker heterozygosity on the dynamics of the Markov approximation in a recombining population. Furthermore, Figures 1 and 4–6 all show that the performance of the Markov approximation is worse at intermediate distances than at both short and long distances. These results are explained in the next section by reference to two contrasting processes of mixing and perturbation.

For all measures, the baseline error measures do not converge to zero at large genomic distances, as would be the case in the absence of linkage disequilibrium. The main reason for this is that our sample size is small— even if a pair of markers is in perfect linkage equilibrium in a population, a small sample from that population will contain some LD due to sampling error. A second reason is that some long-range LD may be present in the population, due, for example, to admixing or preferential mating.

Position profiles: We now assess how the accuracy of the independent and Markov approximations varies along each chromosome in comparison with local recombination rates. Statistics $Y_{d,n}$ and $Z_{d,n}$ and average recombination rates were calculated for a sliding window of 20 Mb across each chromosome. We used fixed

TABLE 3

Correlation between recombination rates and error measures for sliding window over individual chromosomes

	Haplotype blocks		Individual SNPs	
Chromosome	Markov	Independent	Markov	Independent
1	0.519	−0.794	−0.381	−0.594
2	0.531	−0.791	−0.275	−0.719
3	0.618	−0.797	−0.652	−0.774
4	0.665	−0.788	−0.540	−0.735
5	0.474	−0.851	−0.724	−0.855
6	0.611	−0.766	−0.561	−0.865
7	0.132	−0.887	−0.667	−0.882
8	0.803	−0.690	−0.290	−0.775
9	−0.008	−0.675	0.027	−0.531
10	0.522	−0.823	−0.583	−0.812
11	0.468	−0.762	−0.442	−0.660
12	0.826	−0.806	−0.710	−0.811
13	0.816	−0.949	−0.860	−0.932
14	0.455	−0.909	−0.695	−0.914
15	0.833	−0.848	0.119	−0.793
16	0.522	−0.734	0.549	0.002
17	0.241	−0.812	−0.655	−0.626
18	0.571	−0.927	−0.832	−0.853
19	0.595	−0.619	−0.662	−0.826
20	0.941	−0.503	−0.711	−0.904
21	0.060	0.355	0.682	0.640
22	−0.069	−0.142	−0.718	−0.439
1–22 mean	0.506	−0.705	−0.435	−0.666
1–22 SD	0.281	0.292	0.423	0.358
1–20 mean	0.557	−0.787	−0.477	−0.743
1–20 SD	0.238	0.107	0.355	0.207

The haplotype block column refers to the HaploBlock model with up to four variants.

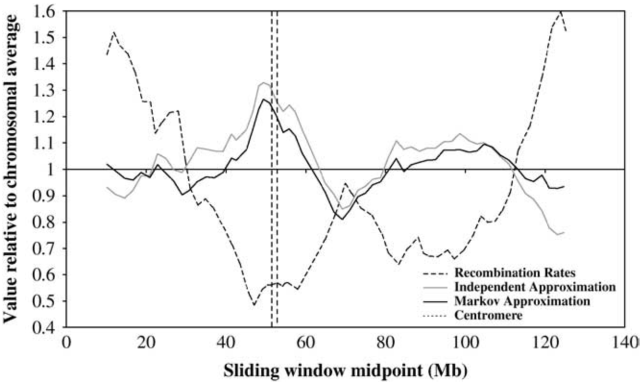

Figure 7.—Position profiles for individual SNP models over chromosome 11.

values of $d = 500$ kb and $n = 4$ for all the analyses. Local recombination rates were taken from the deCODE map and aligned against the genome build for our HapMap data, using the University of California Santa Cruz Table Browser (Kong *et al.* 2002; Karolchik *et al.* 2004).

We correlated the error measures and the recombination rates over the window positions for each chromosome. Table 3 shows the correlation coefficients for the HaploBlock model with up to four variants per block and for individual SNPs. Windows with low SNP density due to their proximity to a centromere were excluded from these calculations. As can be seen in Table 3, the Markov approximation for the HaploBlock model shows a positive correlation between recombination rates and approximation error, with an average coefficient over the chromosomes of 0.506 ± 0.281. This contrasts with the independent approximation for the HaploBlock model, with an average coefficient of −0.705 ± 0.292.

When considering SNPs individually, a different picture emerges. The error rates of both the independent and Markov approximations are lower in regions of high recombination, just as in the independent approximation for the HaploBlock model. For the Markov approximation over individual SNPs, the average coefficient of correlation over the chromosomes is −0.435 ± 0.423. The performance of the independent approximation over individual SNPs is similar to that for the HaploBlock model, with an average coefficient of −0.666 ± 0.358.

The correlation coefficients for chromosomes 21 and 22 in Table 3 differ significantly from the mean values in many cases. This is because the HapMap data cover just 37 Mb of chromosome 21 and 35 Mb of chromosome 22, so that a sliding window of 20 Mb produces a weak signal. Table 3 also shows the results obtained if these chromosomes are removed from the sample, by averaging those for chromosomes 1–20. In all cases, this reduces the standard deviation of the values and strengthens the average correlation.

It is instructive to look at one chromosome in more depth, to see an example of how the error measures vary in comparison to local recombination rates. We examine here chromosome 11, since its correlation coefficients as shown in Table 3 are close to the averages over all of the chromosomes. Figure 7 shows how the individual SNP approximation errors vary with recombination rates over the chromosome. As can be seen, the error rates of the two approximations follow each other closely and are strongly anticorrelated with recombination rates. At the ends of the chromosome where recombination rates are highest, both approximations perform well. At the centromere, where recombination rates are generally lower, the opposite effect is seen. In particular, recombination rates near the centromere are ~50% of the average, while the Markov and independent approximation error is 20–30% higher than the average.

Figure 8 shows the equivalent relationship for the HaploBlock model with up to four variants. The independent approximation performs best at the chromosome ends where recombination rates are highest and worst near the centromere where they are low. The

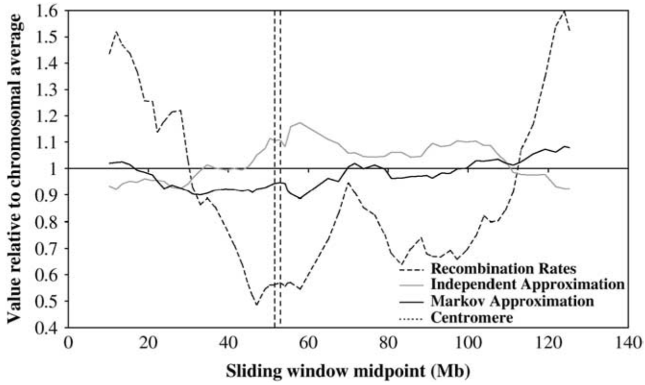

FIGURE 8.—Position profiles for haplotype block models over chromosome 11.

behavior is very similar to that presented in Figure 7 for individual SNPs. By contrast, the Markov approximation for blocks performs worst at the ends of the chromosome and best near the centromere. Consequently, unlike the SNP approximations shown in Figure 7, the independent and Markov approximations for haplotype blocks are significantly out of phase.

Figure 9 summarizes the behavior of the Markov position profiles for all the different models examined. Each point compares the average marker heterozygosity for a particular model against the average gradient for that model of the best-fit line between the Markov error measure and recombination rates. This gradient $\Delta_Z/\Delta_{\mathrm{cM/Mb}}$ measures the strength of the effect of local recombination rates on the local performance of the Markov model. Figure 9 shows that for each set of related models, this strength rises monotonically with the average heterozygosity. In the section that follows, we provide a theoretical explanation for this three-way relationship between heterozygosity, recombination, and the accuracy of the Markov approximation.

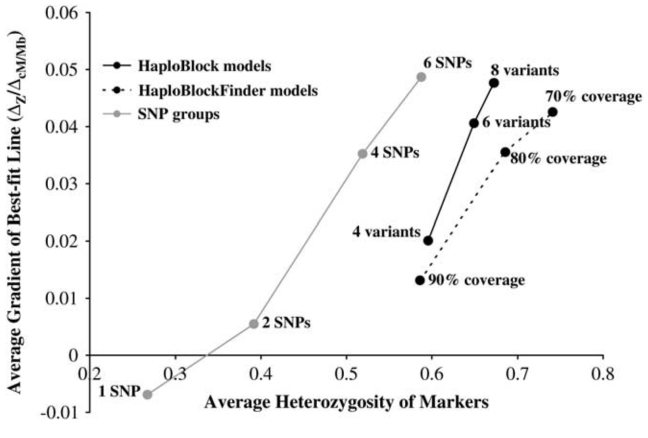

FIGURE 9.—Effect of heterozygosity on average best-fit gradient between Markov error and recombination rates.

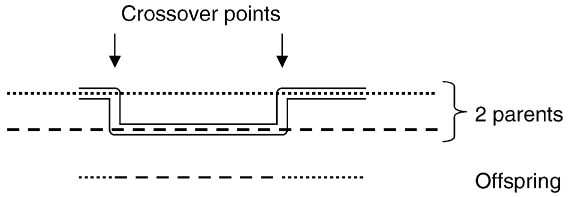

FIGURE 10.—Meiotic recombination.

ON THE MARKOV MODEL

Mixing and perturbation: We define mixing as the progressive reduction in linkage disequilibrium between the markers on a chromosome, as a result of recombination. In a theoretical closed population with random mating, all markers on a chromosome will converge on perfect linkage equilibrium (GEIRINGER 1944). However, the speed of the mixing process depends on two key factors: (a) mixing is faster between more distant markers due to the higher probability of recombination, and (b) mixing is faster between markers with fewer alleles (*e.g.*, SNPs) since each recombination is more likely to bring the marker distribution closer to equilibrium (RABANI *et al.* 1998; ARDLIE *et al.* 2002; VARILO *et al.* 2003). Since the independent approximation error stems from linkage disequilibrium, the speed of mixing also determines the accuracy of this approximation at different distances. An independent model is a special case of a Markov model, so the mixing process also contributes to the accuracy of the Markov approximation.

We introduce here a second process related to recombination called perturbation, which affects only the Markov approximation. Perturbation is defined as the introduction of new long-range correlations between markers on a chromosome, as a result of double recombinations. These long-range correlations contribute to inaccuracy in the Markov model. Let us assume that two parent haplotypes are completely distinct from each other. The joint distribution over any set of markers in the parent haplotypes can be represented perfectly by a Markov model, since the allele at each variable site completely determines that at the next site. However, offspring haplotypes produced by double recombination from these parents receive two disjoint sections from one parent, separated by a section from the other parent, as shown in Figure 10. In these cases, the correlation between the disjoint sections cannot be expressed in terms of the intermediate region. Since the Markov model represents only dependencies between immediately adjacent markers, these double recombinations introduce inaccuracy in the Markov approximation for the offspring that was not present in the parents. As with the mixing process, this perturbation effect is strongest where the probability of recombination is higher, since this also means a higher probability of double recombinations.

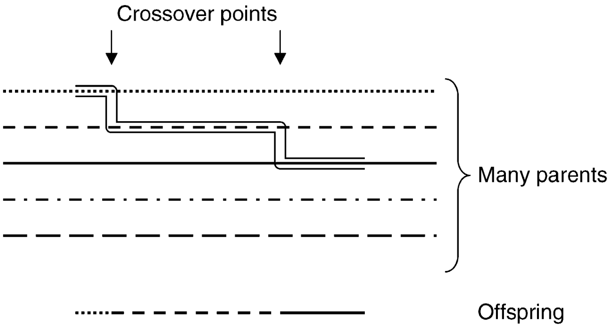

FIGURE 11.—Intermixing.

The perturbation process constitutes a key difference between the dynamics of the independent and Markov models. In an infinite population, the accuracy of the independent approximation for a set of markers increases monotonically from one generation to the next. By contrast, the accuracy of the Markov approximation can increase or decrease, depending on the relative intensity of the mixing and perturbation processes. As we show later, the perturbation process can be stronger for markers with a larger number of alleles, rendering it more visible for multiallelic haplotype blocks than for biallelic SNP markers. This explains why we see a positive correlation between recombination rates and Markov approximation error for blocks, where perturbation is pronounced, but do not see this effect when modeling individual SNPs where perturbation is weaker.

The complex relationship shown in Figures 1 and 4–6 between physical distance and the accuracy of the Markov approximation for haplotype blocks is also explained by the balance between mixing and perturbation. At short distances, the Markov approximation over blocks is accurate due to the low probability of double recombination and the consequent lack of perturbation. At long distances, the Markov approximation over blocks is accurate due to the high probability of recombination and the consequent strong mixing. At intermediate distances, some perturbation takes place but mixing is weak, so the performance of the Markov approximation over haplotype blocks is at its worst.

Intermixing: For meiotic recombination under random mating, an offspring haplotype is generated from two parent haplotypes by the process depicted in Figure 10. Two parent haplotypes are selected independently from the source population. The offspring haplotype is generated from these parents by a reading process that crosses over from one parent to the other with probability θ_j between markers j and $j + 1$, where θ_j is the recombination fraction between the markers. As a result, the offspring haplotype can contain alternating stretches of genetic material from the two parents.

Our proof makes use of a different process called intermixing. Figure 11 depicts the intermixing process with the same crossover points as the meiosis in Figure 10. In intermixing, a large number of parent haplotypes are selected independently from the source population. The offspring haplotype is generated from these parents by a reading process that moves to a *new* parent with probability θ_j between markers j and $j + 1$. An offspring haplotype generated by intermixing with x crossovers will contain genetic material from $x + 1$ independently selected parents. In contrast to normal meiosis, the theoretical intermixing process cannot introduce new long-range dependencies, since the reading process never returns to a parent previously used.

The key point for our purposes is that if the first two intermixing parents are the same as those for meiosis, the results of meiosis and intermixing are identical if no more than one crossover took place. With less than two crossovers, intermixing uses only the first two parent haplotypes, producing the same offspring haplotype as meiosis. Differences arise only due to double crossovers, after which meiosis returns to the first parent haplotype whereas intermixing selects a new parent. The proof that follows is based on this similarity between the two processes and the fact that intermixing preserves the Markov properties of a population regardless of how many crossovers take place.

Theorem: Consider a population of infinite size in Hardy–Weinberg equilibrium. This population undergoes random mating and meiotic recombination without interference in a series of discrete generations. Consider a set of n markers numbered $1, \ldots, n$, with recombination fraction θ_j between each pair of adjacent markers j and $j + 1$.

Define $P_u(x_1, \ldots, x_n)$ as the haplotype distribution over sites $1, \ldots, n$ in generation u and $Q_u(x_1, \ldots, x_n) = P_u(x_1)\prod_{i=1}^{n-1} P_u(x_{i+1} \mid x_i)$ as its Markov approximation. Similarly, P_{u+1} is the haplotype distribution that emerges in generation $u + 1$ and Q_{u+1} is its Markov approximation.

We define $Z_u = \|P_u - Q_u\|$ as the variation distance between distributions P_u and Q_u, and $Z_{u+1} = \|P_{u+1} - Q_{u+1}\|$. Let $D_u(j) = 1 - \sum_{x_j}(P_u(x_j))^2$ be the heterozygosity of site j in generation u, defined by the probability that two haplotypes chosen randomly from distribution P_u differ at site j. Our theorem states that for $n \leq 5$

$$Z_{u+1} \leq Z_u + \frac{1}{2}\left(\sum_{i=1}^{n-1}\theta_i\right)^2 \cdot \min\left(1, \sum_{j=3}^{n} D_u(j)\right). \qquad (1)$$

Thus, the error Z_{u+1} of the Markov approximation in generation $u + 1$ is bounded by the error Z_u in generation u, plus an additional term that depends on two factors. The first factor is the square of the total of the intermarker recombination fractions. The second factor is the sum of the heterozygosities of sites $3, \ldots, n$, bounded to be no more than 1.

A full proof of Equation 1 for $n \leq 5$ is provided in the APPENDIX. The outline is as follows. Let P'_{u+1} be the distribution that emerges from performing intermixing on generation u and Q'_{u+1} be its Markov approximation. We use P'_{u+1} and Q'_{u+1} to prove the bound on $Z_{u+1} = \|P_{u+1} - Q_{u+1}\|$ by applying the triangular inequality

$$\|P_{u+1} - Q_{u+1}\| \leq \|P_{u+1} - P'_{u+1}\| + \|P'_{u+1} - Q'_{u+1}\| + \|Q'_{u+1} - Q_{u+1}\|.$$

The first step is to prove an upper bound on $\|P_{u+1} - P'_{u+1}\|$, the variation distance between the haplotype distributions generated by meiosis and intermixing. This distance is bounded by $\frac{1}{2}\left(\sum_{i=1}^{n-1}\theta_i\right)^2 \cdot \min(1, \sum_{j=3}^{n} D_u(j))$. The intuition here is that the results of meiosis and intermixing differ only if there was a double recombination, the probability of which is bounded by $\left(\sum_{i=1}^{n-1}\theta_i\right)^2$. If a double recombination did occur, the probability that the offspring haplotype will differ between meiosis and intermixing is bounded by the sum of the heterozygosities $D_u(j)$ for sites $j = 3, \ldots, n$, since $j = 3$ is the first site that can be affected by a double recombination.

The second step is to bound $\|P'_{u+1} - Q'_{u+1}\|$, the variation distance between the distribution resulting from intermixing and its Markov approximation. We prove that for $n \leq 5$, this distance is no greater than $\|P_u - Q_u\| = Z_u$. This result arises because each crossover event in intermixing selects a new parent haplotype at random, so no new long-range dependencies are introduced. A proof of this bound for $n \leq 5$ is provided in the APPENDIX. We also conjecture that this bound holds true for all values of n, as suggested by extensive simulation.

The final step is to prove that $\|Q'_{u+1} - Q_{u+1}\| = 0$, namely that the Markov approximations of the distributions arising from meiosis and intermixing are identical. The intuition here is that the Markov approximation is entirely determined by the joint distribution over each pair of adjacent sites, and this joint distribution is identical for both intermixing and meiosis.

These results are combined under the triangular inequality to yield Equation 1:

$$\|P_{u+1} - Q_{u+1}\| \leq \|P_{u+1} - P'_{u+1}\| + \|P'_{u+1} - Q'_{u+1}\| + \|Q'_{u+1} - Q_{u+1}\|$$
$$\leq \frac{1}{2}\left(\sum_{i=1}^{n-1}\theta_i\right)^2 \cdot \min\left(1, \sum_{j=3}^{n} D_u(j)\right) + Z_u.$$

The average heterozygosity for individual SNPs in the HapMap data is 0.267. By contrast, all of the HaploBlock and HaploBlockFinder block models have an average heterozygosity of 0.586 or more, more than double that for individual SNPs (see Figure 9). Equation 1 suggests that increased heterozygosity leads to a stronger perturbation process, which in turn explains the difference in behavior of the Markov approximation for different types of marker. Nonetheless, since Equation 1 provides only an upper bound, it does not provide a complete explanation of this relationship. More theoretical work is required to identify a lower bound, as well as additional factors that affect the perturbation process.

DISCUSSION

In this work, we assessed the accuracy of the independent and Markov approximations for representing background variation in the human genome. Using data taken from HapMap, we showed how the approximation error varies for different physical distances and along each autosome, when modeling both individual SNPs and haplotype blocks of various models. Our core observation is that the Markov model over haplotype blocks is particularly accurate at representing markers in strong linkage disequilibrium. By reference to the perturbation process, we explained why the Markov approximation exhibits this behavior only when modeling multiallelic haplotype blocks, rather than biallelic individual SNPs.

Our motivation for this work was to assess whether it is important to use a Markov chain to represent haplotype block variation or whether an independent model suffices. Clearly, a Markov approximation can represent the variation for a set of markers more accurately than an independent approximation, due to the larger number of parameters available. However, our results show an important additional benefit of the Markov model—that when used with haplotype blocks, it is uniquely suited for modeling genomic variation at high density. Models of background variation combining haplotype blocks and a Markov chain have been used by ourselves and others (DALY *et al.* 2001; ANDERSON and NOVEMBRE 2003; GREENSPAN and GEIGER 2004a; KIMMEL and SHAMIR 2004).

The error measure we employed is based on the variation distance between a joint distribution and its maximum-likelihood approximation. We used this measure because it permits direct comparison between the independent and Markov approximations and has an intuitive interpretation in terms of the proportion of a distribution misrepresented by its approximation. However, this measure is not ideal, since it is biased by the allele frequencies at individual markers, just like the $|D|$ measure of linkage disequilibrium to which it is related. It would be fruitful to develop an equivalent of the D' linkage disequilibrium measure for the Markov model, to overcome this disadvantage. Nonetheless, since our empirical observations were based on averages over large numbers of sites, this shortcoming does not affect the overall patterns observed.

We showed that the unusual accuracy of the Markov model for representing haplotype blocks over short distances stemmed from the fact that blocks have higher heterozygosity than individual SNPs. We also showed that similar results can be achieved by arbitrarily grouping sets of adjacent SNPs into multiallelic markers. This confirms our theoretical result that the behavior of the Markov approximation depends on allele diversity, rather than on a specific model of haplotype blocks. Nonetheless, haplotype blocks based on statistical criteria offer other advantages over arbitrary groups of SNPs in terms of model simplicity and the selection of htSNPs.

We referred above to the dependency of the Markov model on the balance between the mixing and perturbation processes. Beyond our initial observations, there is work to be done in understanding how these two processes interact and in developing more precise criteria for determining when each one plays a more dominant role. It is also desirable to ascertain whether a population must contain highly distinct haplotypes for the perturbation effect to be seen. On this point, recent research has found an abundance of common haplotypes that differ at almost every site in human populations (Zhang *et al.* 2003). Finally, it would be valuable to generalize our proof to a population of finite size and to extend it to more than $n = 5$ sites.

LITERATURE CITED

Anderson, E., and J. Novembre, 2003 Finding haplotype block boundaries by using the minimum-description-length principle. Am. J. Hum. Genet. **73:** 336–354.

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. **3:** 299–309.

Arnheim, N., P. Calabrese and M. Nordborg, 2003 Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. Am. J. Hum. Genet. **73:** 5–16.

Cardon, L., and G. Abecasis, 2003 Using haplotype blocks to map human complex trait loci. Trends Genet. **19:** 135–140.

Daly, M., J. Rioux, S. Schaffner, T. Hudson and E. Lander, 2001 High-resolution haplotype structure in the human genome. Nat. Genet. **29:** 229–232.

Devlin, B., and N. Risch, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics **29:** 311–322.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

Geiringer, H., 1944 On the probability theory of linkage in Mendelian heredity. Ann. Math. Stat. **15:** 25–37.

Goldstein, D., 2001 Islands of linkage disequilibrium. Nat. Genet. **29:** 109–111.

Greenspan, G., and D. Geiger, 2004a Model-based inference of haplotype block variation. J. Comp. Biol. **11:** 493–504.

Greenspan, G., and D. Geiger, 2004b High density linkage disequilibrium mapping using models of haplotype block variation. Bioinformatics **20**(Suppl. 1): I137–I144.

Hardy, G. H., 1908 Mendelian proportions in a mixed population. Science **18:** 49–50.

International HapMap Consortium, 2003 The International HapMap Project. Nature **426:** 789–796.

Jeffreys, A., A. Ritchie and R. Neumann, 2000 High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum. Mol. Genet. **9:** 725–733.

Jeffreys, A., L. Kauppi and R. Neumann, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29:** 217–222.

Johnson, G. C., L. Esposito, B. J. Barratt, A. N. Smith, J. Heward *et al.*, 2001 Haplotype tagging for the identification of common disease genes. Nat. Genet. **29:** 233–237.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet *et al.*, 2004 The UCSC Table Browser data retrieval tool. Nucleic Acids Res. **32:** D493–D496.

Kimmel, G., and R. Shamir, 2004 Maximum likelihood resolution of multi-block genotypes. Proceedings of the Eighth Annual International Conference on Computational Molecular Biology (RECOMB 2004), San Diego, March 27–31, pp. 2–9.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Liu, J., C. Sabatti, J. Teng, B. Keats and N. Risch, 2001 Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. **11:** 1716–1724.

McPeek, M., and A. Strahs, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am. J. Hum. Genet. **65:** 858–875.

Morris, A., J. Whittaker and D. Balding, 2000 Bayesian fine-scale mapping of disease loci, by hidden Markov models. Am. J. Hum. Genet. **67:** 155–169.

Morris, A. P., J. C. Whittaker and D. J. Balding, 2002 Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am. J. Hum. Genet. **70:** 686–707.

Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294:** 1719–1723.

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding *et al.*, 2003 Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat. Genet. **33:** 382–387.

Rabani, Y., Y. Rabinovich and A. Sinclair, 1998 A computational view of population genetics. Random Struct. Algorithms **12:** 313–334.

Sebastiani, P., R. Lazarus, S. T. Weiss, L. M. Kunkel, I. S. Kohane *et al.*, 2003 Minimal haplotype tagging. Proc. Natl. Acad. Sci. USA **100:** 9900–9905.

Templeton, A., A. Clark, K. Weiss, D. Nickerson, E. Boerwinkle *et al.*, 2000 Recombinational and mutational hotspots within the human lipoprotein lipase gene. Am. J. Hum. Genet. **66:** 69–83.

Tishkoff, S. A., and B. C. Verrelli, 2003 Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. Curr. Opin. Genet. Dev. **13:** 569–575.

Twells, R. C., C. A. Mein, M. S. Phillips, J. F. Hess, R. Veijola *et al.*, 2003 Haplotype structure, LD blocks, and uneven recombination within the LRP5 gene. Genome Res. **13:** 845–855.

Varilo, T., T. Paunio, A. Parker, M. Perola, J. Meyer *et al.*, 2003 The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. Hum. Mol. Genet. **12:** 51–59.

Wall, J. D., and J. K. Pritchard, 2003 Assessing the performance of the haplotype block model of linkage disequilibrium. Am. J. Hum. Genet. **73:** 502–515.

Wang, N., J. Akey, K. Zhang, R. Chakraborty and L. Jin, 2002 Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am. J. Hum. Genet. **71:** 1227–1234.

Zhang, K., and L. Jin, 2003 HaploBlockFinder: haplotype block analyses. Bioinformatics **19:** 1300–1301.

Zhang, J., W. L. Rowe, A. G. Clark and K. H. Buetow, 2003 Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. Am. J. Hum. Genet. **73:** 1073–1081.

Zhang, K., P. Calabrese, M. Nordborg and F. Sun, 2002 Haplotype block structure and its applications to association studies: power and study designs. Am. J. Hum. Genet. **71:** 1386–1394.

APPENDIX

Here we prove in full the theoretical result outlined in this article.

Definitions: Under meiotic recombination, each offspring haplotype over n sites is formed from two parent haplotypes $y^1 = (y_1^1, \ldots, y_n^1)$ and $y^2 = (y_1^2, \ldots, y_n^2)$. Each meiosis entails a crossover vector $r = (r_1, \ldots, r_n) \in \{0, 1\}^{n-1}$, in which $r_i = 1$ if a crossover took place between sites i and $i + 1$ and $r_i = 0$ otherwise. Let $F(y^1, y^2, r)$ denote the offspring haplotype that is generated by meiosis from y^1 and y^2, assuming a crossover vector r:

$$F(y^1, y^2, r) = y_1^{S(r,1)}, \ldots, y_n^{S(r,n)}. \tag{A1}$$

In Equation A1, $S(r; i)$ is the index of the parent of site i in the offspring, namely $S(r, i) = 1 + \sum_{k=1}^{i-1} r_k$ modulo 2. If there are an even number of recombinations up to site i then $S(r; i) = 1$; otherwise $S(r; i) = 2$. Since both parents are selected randomly from the same distribution, we assumed without loss of generality that the first site in the offspring comes from parent haplotype y^1.

The probability of a crossover occurring between sites i and $i + 1$ is denoted by θ_i. We define the probability $G(r)$ of a crossover vector r in terms of these pairwise probabilities:

$$G(r_1, \ldots, r_{n-1}) = \prod_{i=1}^{n-1} \theta_i^{r_i} \cdot (1 - \theta_i)^{1-r_i}. \tag{A2}$$

Recall that $P_u(x)$ denotes the frequency of haplotype x in generation u. The frequency $P_{u+1}(x)$ of haplotype x in generation $u + 1$ due to meiotic recombination is the sum of the probabilities of all joint assignments to y^1, y^2, and r, which yield x:

$$P_{u+1}(x) = \sum_{y^1, y^2, r | F(y^1, y^2, r) = x} G(r) P_u(y^1) P_u(y^2). \tag{A3}$$

For intermixing over n sites, each offspring haplotype can inherit sections from up to n haplotypes in the previous generation, although in most cases less than n will be used. Let $F'(y^1, \ldots, y^n, r)$ denote the haplotype generated from $y^1, \ldots, y^n$ by intermixing under a crossover vector r:

$$F'(y^1, y^2, r) = y_1^{S'(r,1)}, \ldots, y_n^{S'(r,n)}. \tag{A4}$$

In Equation A4, $S'(r; i)$ is the index of the parent of site i in the offspring, namely $S'(r, i) = 1 + \sum_{k=1}^{i-1} r_k$. The function $S'(r; i)$ counts the number of crossovers that have taken place up to site i. The frequency $P'_{u+1}(x)$ of haplotype x in generation $u + 1$ due to intermixing on parent distribution P_u is as follows:

$$P'_{u+1}(x) = \sum_{y^1, \ldots, y^n, r | F'(y^1, \ldots, y^n, r) = x} G(r) \prod_{i=1}^{n} P_u(y^i). \tag{A5}$$

Intermixing and meiosis: We prove the following bound on the variation distance between the haplotype distribution P_{u+1} arising from meiosis on generation u and the distribution P'_{u+1} arising from intermixing:

$$\|P_{u+1} - P'_{u+1}\| \leq \frac{1}{2} \left(\sum_{i=1}^{n-1} \theta_i \right)^2 \cdot \min\left(1, \sum_{j=3}^{n} D_u(j) \right). \tag{A6}$$

Recall that $D_u(j)$ is defined as the heterozygosity of site j in generation u, where $D_u(j) = 1 - \sum_{x_j} (P_u(x_j))^2$ is the probability that two haplotypes randomly chosen from P_u differ at site j.

Let $R = \{0, 1\}^{n-1}$ denote the set of all possible crossover vectors r. Let R^- be the subset $\{r \in R \,|\, \sum_j r_j \leq 1\}$ consisting of crossover vectors representing ≤ 1 crossovers, and let $R^+ = \{r \in R \,|\, \sum_j r_j \geq 2\}$ denote the subset representing ≥ 2 crossovers. Clearly, $R = R^- \cup R^+$ and $R^- \cap R^+ = \emptyset$. The frequency of haplotype x after meiosis, given in Equation A3, can therefore be written as

$$P_{u+1}(x) = \sum_{y^1, y^2, r \in R^- | F(y^1, y^2, r) = x} G(r) P_u(y^1) P_u(y^2) + \sum_{y^1, y^2, r \in R^+ | F(y^1, y^2, r) = x} G(r) P_u(y^1) P_u(y^2). \tag{A7}$$

Similarly, the frequency of x after intermixing, given in Equation A5, can be written as

$$P'_{u+1}(x) = \sum_{y^1, \ldots, y^n, r \in R^- | F'(y^1, \ldots, y^n, r) = x} G(r) \prod_{i=1}^{n} P_u(y^i) + \sum_{y^1, \ldots, y^n, r \in R^+ | F'(y^1, \ldots, y^n, r) = x} G(r) \prod_{i=1}^{n} P_u(y^i). \tag{A8}$$

Recall that if $r \in R^-$ then $\sum_j r_j \leq 1$. In these cases, $S(r, i) = S'(r, i)$ for all i, yielding $F'(y^1, \ldots, y^n, r) = F(y^1, y^2, r)$. In other words, when less than two crossovers occur, the haplotype obtained by meiosis is identical to that obtained by intermixing for the same parents y^1 and y^2. Consequently, we rewrite Equation A8 as follows:

$$P'_{u+1}(x) = \sum_{y^1,y^2,r\in R^-|F(y^1,y^2,r)=x} G(r)P_u(y^1)P_u(y^2) + \sum_{y^1,\ldots,y^n,r\in R^+|F'(y^1,\ldots,y^n,r)=x} G(r)\prod_{i=1}^{n} P_u(y^i). \tag{A9}$$

Since the sums in Equations A7 and A9 corresponding to no more than one crossover are identical, the variation distance between P_{u+1} and P'_{u+1} is due to two or more crossovers:

$$\|P_{u+1} - P'_{u+1}\| = \frac{1}{2}\sum_x \left| \sum_{y^1,y^2,r\in R^+|F(y^1,y^2,r)=x} G(r)P_u(y^1)P_u(y^2) - \sum_{y^1,\ldots,y^n,r\in R^+|F'(y^1,\ldots,y^n,r)=x} G(r)\prod_{i=1}^{n} P_u(y^i) \right|. \tag{A10}$$

By introducing the unity sum $\sum_{y^3,\ldots,y^n}\prod_{i=3}^{n} P_u(y^i) = 1$ into the first term of Equation A10, we obtain

$$\|P_{u+1} - P'_{u+1}\| = \frac{1}{2}\sum_x \left| \sum_{y^1,\ldots,y^n,r\in R^+|F(y^1,y^2,r)=x} G(r)\prod_{i=1}^{n} P_u(y^i) - \sum_{y^1,\ldots,y^n,r\in R^+|F'(y^1,\ldots,y^n,r)=x} G(r)\prod_{i=1}^{n} P_u(y^i) \right|. \tag{A11}$$

We now derive the bound for $\|P_{u+1} - P'_{u+1}\|$, as given by Equation A6. Let $[a = b]$ denote the function that returns 1 if $a = b$ and 0 otherwise, and define $[a \neq b] = 1 - [a = b]$. Equation A11 is reformulated as follows:

$$\begin{aligned} \|P_{u+1} - P'_{u+1}\| &= \frac{1}{2}\sum_x \left| \sum_{y^1,\ldots,y^n,r\in R^+} G(r)\prod_{i=1}^{n} P_u(y^i)\cdot\{[F(y^1,y^2,r) = x] - [F'(y^1,\ldots,y^n,r) = x]\} \right| \\ &\leq \sum_{r\in R^+} G(r) \sum_{y^1,\ldots,y^n}\prod_{i=1}^{n} P_u(y^i)\cdot\frac{1}{2}\sum_x |[F(y^1,y^2,r) = x] - [F'(y^1,\ldots,y^n,r) = x]| \\ &= \sum_{r\in R^+} G(r) \sum_{y^1,\ldots,y^n}\prod_{i=1}^{n} P_u(y^i)\cdot[F(y^1,y^2,r) \neq F'(y^1,\ldots,y^n,r)]. \end{aligned} \tag{A12}$$

The last equality follows because if $F(y^1, y^2, r) = F'(y^1, \ldots, y^n, r)$ then the expression $|[F(y^1, y^2, r) = x] - [F'(y^1, \ldots, y^n, r) = x]| = 0$ for all x, and if $F(y^1, y^2, r) \neq F'(y^1, \ldots, y^n, r)$, then $|[F(y^1, y^2, r) = x] - [F'(y^1, \ldots, y^n, r) = x]| = 1$ for exactly two values of x, namely $x = F(y^1, y^2, r)$ and $x = F'(y^1, \ldots, y^n, r)$.

The value $[F(y^1, y^2, r) \neq F'(y^1, \ldots, y^n, r)] = 1$ if the haplotype that arises from meiosis is different from that arising from intermixing. This condition is fulfilled if the haplotypes differ in at least one site. The haplotypes are always identical at sites 1 and 2 since the earliest that an observed double recombination can occur is between sites 2 and 3. In other words, $S(r, 1) = S'(r, 1)$ and $S(r, 2) = S'(r, 2)$ for any crossover vector r. By summing the possibilities for the remaining sites 3, $\ldots$, n, we obtain a simple bound:

$$[F(y^1, y^2, r) \neq F'(y^1, \ldots, y^n, r)] \leq \sum_{j=3}^{n}[y_j^{S(r,j)} \neq y_j^{S'(r,j)}]. \tag{A13}$$

Equations A12 and A13 yield

$$\|P_{u+1} - P'_{u+1}\| \leq \sum_{r\in R^+} G(r)\sum_{j=3}^{n}\sum_{y^1,\ldots,y^n}\prod_{i=1}^{n} P_u(y^i)\cdot[y_j^{S(r,j)} \neq y_j^{S'(r,j)}]. \tag{A14}$$

Since, in the worst case, every site from the third one onward has a different source under meiosis and intermixing, $\sum_{y^1,\ldots,y^n}\prod_{i=1}^{n} P_u(y^i)\cdot[y_j^{S(r,j)} \neq y_j^{S'(r,j)}]$ is the probability that two independently selected haplotypes from distribution P_u differ at site j. This is precisely the definition of heterozygosity $D_u(j)$, so

$$\|P_{u+1} - P'_{u+1}\| \leq \sum_{r\in R^+} G(r)\sum_{j=3}^{n} D_u(j). \tag{A15}$$

Since $[F(y^1, y^2, r) \neq F'(y^1, \ldots, y^n, r)] \leq 1$ by definition, an additional bound is obtained for $\|P_{u+1} - P'_{u+1}\|$ from Equation A12:

$$\|P_{u+1} - P'_{u+1}\| \leq \sum_{r\in R^+} G(r)\sum_{y^1,\ldots,y^n}\prod_{i=1}^{n} P_u(y^i) = \sum_{r\in R^+} G(r). \tag{A16}$$

Finally, using the probability $G(r)$ of a crossover vector r (Equation A2), we bound $\sum_{r\in R^+} G(r)$ by summing the probability of every possible pair of crossovers:

$$\sum_{r\in R^+} G(r) \leq \sum_{i=1}^{n-1} \theta_i \sum_{k=i+1}^{n-1} \theta_k \leq \frac{1}{2}\left(\sum_{i=1}^{n-1} \theta_i\right)^2. \tag{A17}$$

Equations A15–A17 yield the bound for $\|P_{u+1} - P'_{u+1}\|$, given by Equation A6.

Markov accuracy after intermixing: Recall that $P'_{u+1}(x)$ is the haplotype distribution that results from intermixing parent haplotype distribution P_u and that $Q'_{u+1}(x)$ is the Markov approximation of $P'_{u+1}(x)$. We prove that for $n \leq 5$

$$\|P'_{u+1} - Q'_{u+1}\| \leq \|P_u - Q_u\|. \tag{A18}$$

For haplotypes with $n > 5$ sites, this problem remains open. However, we conjecture that it is true for all values of n, as confirmed by extensive simulation studies up to $n = 16$.

The formula for $P'_{u+1}(x)$ in Equation A5 is now rewritten in terms of contiguous sections inherited from a parent, using the probability $G(r)$ of each crossover vector r and the probability of the parent haplotype sections that lead to x under r,

$$P'_{u+1}(x) = \sum_{r\in R} G(r) \prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)}),$$

where

$$x_{(r,k)} = x_{L(r,k)}, \ldots, x_{U(r,k)}$$

$$L(r,k) = \min\{i \mid S'(r,i) = k\}$$

$$U(r,k) = \max\{i \mid S'(r,i) = k\}. \tag{A19}$$

In Equation A19, the functions $L(r, k)$ and $U(r, k)$ denote, respectively, the first and last sites in the offspring haplotype that originate from parent $S'(r, i) = k$ under crossover vector r. Recall that $S'(r, i)$ is the index of the parent haplotype for site i of the offspring haplotype when intermixing with crossover vector r. The term $P_u(x_{(r,k)})$ denotes the marginal distribution $P_u(x_{L(r,k)}, \ldots, x_{U(r,k)}) = \sum_{x_1,\ldots,x_{L(r,k)-1},x_{U(r,k)+1},\ldots,x_l} P_u(x_1, \ldots, x_l)$.

The process of intermixing can be viewed as the transformation of a parent haplotype distribution P_u into an offspring distribution P'_{u+1}. This transformation can be decomposed into a series of atomic transformations, one over each possible crossover point. Let P'^{i}_{u+1} be the haplotype distribution obtained from intermixing if crossovers are allowed only over sites 1 to i. In other words, P'^{i}_{u+1} is the result of intermixing on P_u if all values $\theta_i, \ldots, \theta_{n-1}$ are set to zero. Clearly, the distribution P'^{1}_{u+1} equals the parent haplotype distribution P_u, since P'^{1}_{u+1} is the result of intermixing if no crossing over is allowed. Similarly, the distribution P'^{n}_{u+1} equals the distribution P'_{u+1} that emerges from intermixing over all sites, since the full set of crossovers between sites 1 and n is allowed. As a result, the transformation $P_u \rightarrow P'_{u+1}$ can be expressed as a series of transformations $P'^{1}_{u+1} \rightarrow P'^{2}_{u+1} \rightarrow \cdots \rightarrow P'^{n}_{u+1}$, where each step $P'^{i}_{u+1} \rightarrow P'^{i+1}_{u+1}$ in the series introduces an additional crossover point between sites i and $i + 1$.

Let R^i be the set of crossover vectors in which crossovers occur only between sites 1 to i; *i.e.*, $R^i = \{r \in R \mid r_i = 0, \ldots, r_{n-1} = 0\}$. Let $G^i(r)$ be the probability of crossover vector $r \in R^i$, defined as follows:

$$G^i(r_1, \ldots, r_{n-1}) = \prod_{j=1}^{i-1} \theta_j^{r_j} \cdot (1 - \theta_j)^{1-r_j}.$$

Using these definitions, the probability $P'^{i}_{u+1}(x)$ of haplotype x after intermixing over sites $1, \ldots, i$ is analogous to $P'_{u+1}(x)$, given in Equation A19:

$$P'^{i}_{u+1}(x_1, \ldots, x_n) = \sum_{r\in R^i} G^i(r) \prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)}) = \sum_{r\in R^i} G^i(r)\left(\prod_{k=1}^{S'(r,n)-1} P_u(x_{(r,k)})\right) P_u(x_{L(r,S'(r,n))}, \ldots, x_n). \tag{A20}$$

The recurrence relation between P'^{i+1}_{u+1} and P'^{i}_{u+1} is explicated by splitting $P'^{i+1}_{u+1}(x)$ into two parts:

$$\begin{aligned}
P'^{i+1}_{u+1}(x) &= \sum_{r\in R^{i+1}|r_i=0} G^{i+1}(r)\prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)}) + \sum_{r\in R^{i+1}|r_i=1} G^{i+1}(r)\prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)})\\
&= (1-\theta_i)\sum_{r\in R^{i+1}|r_i=0} G^{i}(r)\prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)})\\
&\quad + \theta_i \sum_{r\in R^{i+1}|r_i=1} G^{i}(r)\left(\prod_{k=1}^{S'(r,n)-1} P_u(x_{(r,k)})\right) P_u(x_{L(r,S'(r,n))},\ldots,x_n).
\end{aligned}$$

If $r_i = 0$ then no recombination took place between sites i and $i+1$, so the sum over $r \in R^{i+1}$ is the same as that over $r \in R^i$. If $r_i = 1$ then the last recombination took place between i and $i+1$, so $U(r, S'(r, n) - 1) = i$ and $L(r, S'(r, n)) = i + 1$. Consequently,

$$\begin{aligned}
P'^{i+1}_{u+1}(x) &= (1-\theta_i)\sum_{r\in R^{i}} G^{i}(r)\prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)})\\
&\quad + \theta_i \sum_{r\in R^{i+1}|r_i=1} G^{i}(r)\left(\prod_{k=1}^{S'(r,n)-1} P_u(x_{(r,k)})\right) P_u(x_{L(r,S'(r,n))},\ldots,x_n)\\
&= (1-\theta_i)P'^{i}_{u+1}(x_1,\ldots,x_n)\\
&\quad + \theta_i \sum_{r\in R^{i+1}|r_i=1} G^{i}(r)\left(\prod_{k=1}^{S'(r,n)-2} P_u(x_{(r,k)})\right) P_u(x_{L(r,S'(r,n)-1)},\ldots,x_i)P_u(x_{i+1},\ldots,x_n). \qquad \text{(A21)}
\end{aligned}$$

We now replace the sum over $r \in R^{i+1} \mid r_i = 1$ by a different sum over $r' \in R^i$, where each vector r' corresponds to a vector r without the crossover between sites i and $i + 1$:

$$\begin{aligned}
P'^{i+1}_{u+1}(x) &= (1-\theta_i)P'^{i}_{u+1}(x_1,\ldots,x_n)\\
&\quad + \theta_i \sum_{r'\in R^{i}} G^{i}(r')\left(\prod_{k=1}^{S'(r',n)-1} P_u(x_{(r',k)})\right) P_u(x_{L(r',S'(r',n))},\ldots,x_i)P_u(x_{i+1},\ldots,x_n)\\
&= (1-\theta_i)\cdot P'^{i}_{u+1}(x_1,\ldots,x_n) + \theta_i\cdot P'^{i}_{u+1}(x_1,\ldots,x_i)P_u(x_{i+1},\ldots,x_n). \qquad \text{(A22)}
\end{aligned}$$

We have replaced $G^i(r)$ with $G^i(r')$ in the transformation from Equation A21 to Equation A22 since the function G^i is not affected by crossovers after site i. The function $S'(r, n)$ in Equation A21 counts the total number of crossovers represented by vector r. It is replaced by $S'(r', n) + 1$ in Equation A22 since r' has one fewer crossover than r. The product of marginal distributions $\prod_{k=1}^{S'(r,n)-2} P_u(x_{(r,k)})$ in Equation A21 is replaced by the product $\prod_{k=1}^{S'(r',n)-1} P_u(x_{(r',k)})$ in Equation A22 since it is related only to chromosomal sections preceding site i, whose parent haplotypes are identical under r and r'. Similarly, $L(r, S'(r, n) - 1)$ in Equation A21 is replaced with $L(r', S'(r', n))$ in Equation A22 since the left edge of the penultimate contiguous section in r that ends at site i becomes the left edge of the last contiguous section in r'.

The distribution P'^{i}_{u+1} is the result of intermixing only up to site i, so its marginal $P'^{i}_{u+1}(x_{i+1},\ldots,x_n)$ over sites $i + 1, \ldots, n$ is the same as the parent marginal $P_u(x_{i+1}, \ldots, x_n)$. Consequently, Equation A22 implies that

$$P'^{i+1}_{u+1}(x) = (1-\theta_i)\cdot P'^{i}_{u+1}(x_1,\ldots,x_n) + \theta_i\cdot P'^{i}_{u+1}(x_1,\ldots,x_i)P'^{i}_{u+1}(x_{i+1},\ldots,x_n). \qquad \text{(A23)}$$

Equation A23 states that the effect of introducing the additional crossover point between sites i and $i + 1$ is to reconstitute a proportion θ_i of the population from the marginal distributions on either side of the crossover point, leaving the remaining $1 - \theta_i$ proportion untouched. Equation A23 also holds in the following marginal form by summing over $x_1, \ldots, x_{i-1}, x_{i+2}, \ldots, x_n$:

$$P'^{i+1}_{u+1}(x_i, x_i{+}1) = (1-\theta_i)\cdot P'^{i}_{u+1}(x_i, x_i{+}1) + \theta_i\cdot P'^{i}_{u+1}(x_i)P'^{i}_{u+1}(x_{i+1}).$$

We now show a similar result for the Markov approximation Q'^{i}_{u+1}, defined as follows:

$$Q'^{i}_{u+1}(x_1, \ldots, x_n) = P'^{i}_{u+1}(x_1) \prod_{j=1}^{n-1} P'^{i}_{u+1}(x_{j+1} \mid x_j). \tag{A24}$$

The recurrence relation between Q'^{i+1}_{u+1} and Q'^{i}_{u+1} is explicated as follows:

$$\begin{aligned}
Q'^{i+1}_{u+1}(x) &= P'^{i+1}_{u+1}(x_1) \prod_{j=1}^{n-1} P'^{i+1}_{u+1}(x_{j+1} \mid x_j) \\
&= P'^{i}_{u+1}(x_1) \prod_{j=1}^{i-1} P'^{i}_{u+1}(x_{j+1} \mid x_j) \cdot P'^{i+1}_{u+1}(x_{i+1} \mid x_i) \cdot \prod_{j=i+1}^{n-1} P'^{i}_{u+1}(x_{j+1} \mid x_j) \\
&= Q'^{i}_{u+1}(x_1, \ldots, x_i) \cdot \frac{P'^{i+1}_{u+1}(x_i, x_{i+1})}{P'^{i+1}_{u+1}(x_i)} \cdot \prod_{j=i+1}^{n-1} P'^{i}_{u+1}(x_{j+1} \mid x_j) \\
&= Q'^{i}_{u+1}(x_1, \ldots, x_i) \cdot \frac{(1-\theta_i)P'^{i}_{u+1}(x_i, x_{i+1}) + \theta_i P'^{i}_{u+1}(x_i) P'^{i}_{u+1}(x_{i+1})}{P'^{i}_{u+1}(x_i)} \cdot \prod_{j=i+1}^{n-1} P'^{i}_{u+1}(x_{j+1} \mid x_j) \\
&= (1-\theta_i) \cdot Q'^{i}_{u+1}(x_1, \ldots, x_i) \cdot P'^{i}_{u+1}(x_{i+1} \mid x_i) \cdot \prod_{j=i+1}^{n-1} P'^{i}_{u+1}(x_{j+1} \mid x_j) \\
&\quad + \theta_i \cdot Q'^{i}_{u+1}(x_1, \ldots, x_i) \cdot P'^{i}_{u+1}(x_{i+1}) \cdot \prod_{j=i+1}^{n-1} P'^{i}_{u+1}(x_{j+1} \mid x_j)
\end{aligned}$$

$$Q'^{i+1}_{u+1}(x) = (1-\theta_i) \cdot Q'^{i}_{u+1}(x_1, \ldots, x_n) + \theta_i \cdot Q'^{i}_{u+1}(x_1, \ldots, x_i) \cdot Q'^{i}_{u+1}(x_{i+1}, \ldots, x_n). \tag{A25}$$

We replaced $P'^{i+1}_{u+1}(x_i)$ with $P'^{i}_{u+1}(x_i)$ at several points above since the intermixing process does not affect the marginal allele frequencies for any individual site. Similarly, we replaced $P'^{i+1}_{u+1}(x_{j+1} \mid x_j)$ with $P'^{i}_{u+1}(x_{j+1} \mid x_j)$ for any $j \neq i$ since the additional crossover permitted between sites i and $i+1$ affects only marginal distributions containing both x_i and x_{i+1}. Equation A25 states the analogous result for the series of Markov approximations $Q'^{1}_{u+1}, \ldots, Q'^{n}_{u+1}$ as Equation A23 states for the series of distributions $P'^{1}_{u+1}, \ldots, P'^{n}_{u+1}$.

Recall that we aim to prove $\|P'_{u+1} - Q'_{u+1}\| \leq \|P_u - Q_u\|$ for $n \leq 5$. Since $P'^{1}_{u+1} = P_u$ and $P'^{n}_{u+1} = P'_{u+1}$, this inequality can be expressed as $\|P'^{n}_{u+1} - Q'^{n}_{u+1}\| \leq \|P'^{1}_{u+1} - Q'^{1}_{u+1}\|$. To establish this inequality, we prove that for $1 \leq i \leq n-1$,

$$\|P'^{i+1}_{u+1} - Q'^{i+1}_{u+1}\| \leq \|P'^{i}_{u+1} - Q'^{i}_{u+1}\|. \tag{A26}$$

We split the proof of Equation A26 into two cases, $i = 1$ and $i = 2$. By considering the haplotypes from their other end points, these proofs also apply, respectively, for $i = n-1$ and $i = n-2$, due to symmetry. This covers all values of $1 \leq i \leq n-1$ provided $n \leq 5$.

Two properties of variation distance are needed. Given two multivariate distributions $A(x, y)$ and $B(x, y)$ with marginal distributions $A(x) = \sum_y A(x, y)$ and $B(x) = \sum_y B(x, y)$, the first property states that $\|A(x, y) - B(x, y)\| \geq \|A(x) - B(x)\|$. Given two mixture distributions $A(x) = \alpha A_1(x) + (1-\alpha)A_2(x)$ and $B(x) = \alpha B_1(x) + (1-\alpha)B_2(x)$, the second property states that $\|A(x) - B(x)\| \leq \alpha\|A_1(x) - B_1(x)\| + (1-\alpha)\|A_2(x) - B_2(x)\|$. Proofs of these two properties are provided at the end of the APPENDIX.

For $i = 1$, we prove Equation A26 by rewriting P'^{2}_{u+1} and Q'^{2}_{u+1} in terms of P'^{1}_{u+1} and Q'^{1}_{u+1}, using the recurrence relations in Equations A23 and A25:

$$P'^{2}_{u+1}(x) = (1-\theta_1) \cdot P'^{1}_{u+1}(x_1, \ldots, x_n) + \theta_1 \cdot P'^{1}_{u+1}(x_1) \cdot P'^{1}_{u+1}(x_2, \ldots, x_n)$$

$$\begin{aligned}
Q'^{2}_{u+1}(x) &= (1-\theta_1) \cdot Q'^{1}_{u+1}(x_1, \ldots, x_n) + \theta_1 \cdot Q'^{1}_{u+1}(x_1) \cdot Q'^{1}_{u+1}(x_2, \ldots, x_n) \\
&= (1-\theta_1) \cdot Q'^{1}_{u+1}(x_1, \ldots, x_n) + \theta_1 \cdot P'^{1}_{u+1}(x_1) \cdot Q'^{1}_{u+1}(x_2, \ldots, x_n).
\end{aligned}$$

The last equality follows because the marginal distribution for an individual site is identical for both P'^{1}_{u+1} and its Markov approximation Q'^{1}_{u+1}. The proof of Equation A26 for $i = 1$ is completed using the two properties of variation distance:

$$\begin{aligned}\|P'^{2}_{u+1} - Q'^{2}_{u+1}\| &\leq (1-\theta_1)\cdot\|P'^{1}_{u+1} - Q'^{1}_{u+1}\| \\ &\quad + \theta_1\cdot\frac{1}{2}\sum_{x_1,\ldots,x_n}|P'^{1}_{u+1}(x_1)P'^{1}_{u+1}(x_2,\ldots,x_n) - P'^{1}_{u+1}(x_1)Q'^{1}_{u+1}(x_2,\ldots,x_n)| \\ &= (1-\theta_1)\cdot\|P'^{1}_{u+1} - Q'^{1}_{u+1}\| \\ &\quad + \theta_1\cdot\frac{1}{2}\sum_{x_1}P'^{1}_{u+1}(x_1)\sum_{x_2,\ldots,x_n}|\,P'^{1}_{u+1}(x_2,\ldots,x_n) - Q'^{1}_{u+1}(x_2,\ldots,x_n)| \\ &\leq (1-\theta_1)\cdot\|P'^{1}_{u+1} - Q'^{1}_{u+1}\| + \theta_1\cdot\|P'^{1}_{u+1} - Q'^{1}_{u+1}\| \\ &= \|P'^{1}_{u+1} - Q'^{1}_{u+1}\|.\end{aligned}$$

For $i = 2$, the proof of Equation A26 proceeds similarly:

$$\begin{aligned}P'^{3}_{u+1}(x) &= (1-\theta_2)\cdot P'^{2}_{u+1}(x_1,\ldots,x_n) + \theta_2\cdot P'^{2}_{u+1}(x_1,x_2)\cdot P'^{2}_{u+1}(x_3,\ldots,x_n) \\ Q'^{3}_{u+1}(x) &= (1-\theta_2)\cdot Q'^{2}_{u+1}(x_1,\ldots,x_n) + \theta_2\cdot Q'^{2}_{u+1}(x_1,x_2)\cdot Q'^{2}_{u+1}(x_3,\ldots,x_n) \\ &= (1-\theta_2)\cdot Q'^{2}_{u+1}(x_1,\ldots,x_n) + \theta_2\cdot P'^{2}_{u+1}(x_1,x_2)\cdot Q'^{2}_{u+1}(x_3,\ldots,x_n).\end{aligned} \tag{A27}$$

The last equality follows since the joint distribution over any two adjacent sites is unchanged by the Markov approximation. The proof of Equation A26 for $i = 2$ is completed using the two properties of variation distance:

$$\begin{aligned}\|P'^{3}_{u+1} - Q'^{3}_{u+1}\| &\leq (1-\theta_2)\cdot\|P'^{2}_{u+1} - Q'^{2}_{u+1}\| \\ &\quad + \theta_2\cdot\frac{1}{2}\sum_{x_1,\ldots,x_n}|P'^{2}_{u+1}(x_1,x_2)P'^{2}_{u+1}(x_3,\ldots,x_n) - P'^{2}_{u+1}(x_1,x_2)Q'^{2}_{u+1}(x_3,\ldots,x_n)| \\ &= (1-\theta_2)\cdot\|P'^{2}_{u+1} - Q'^{2}_{u+1}\| \\ &\quad + \theta_2\cdot\frac{1}{2}\sum_{x_1,x_2}P'^{2}_{u+1}(x_1,x_2)\sum_{x_3,\ldots,x_n}|P'^{2}_{u+1}(x_3,\ldots,x_n) - Q'^{2}_{u+1}(x_3,\ldots,x_n)| \\ &\leq (1-\theta_2)\cdot\|P'^{2}_{u+1} - Q'^{2}_{u+1}\| + \theta_2\cdot\|P'^{2}_{u+1} - Q'^{2}_{u+1}\| \\ &= \|P'^{2}_{u+1} - Q'^{2}_{u+1}\|.\end{aligned} \tag{A28}$$

The proofs for $i = n - 1$ and $i = n - 2$ are obtained by reversing the order of the conditional probabilities in the Markov chain. Since this covers all possible values of $1 \leq i \leq n - 1$ provided $n \leq 5$, this establishes the inequality $\|P'^{i+1}_{u+1} - Q'^{i+1}_{u+1}\| \leq \|P'^{i}_{u+1} - Q'^{i}_{u+1}\|$ and therefore that $\|P'_{u+1} - Q'_{u+1}\| \leq \|P_u - Q_u\|$ for $n \leq 5$, as stated in Equation A18.

For $n > 5$, this method breaks down in Equation A27 for $i = 3$ since the marginal distribution $Q'^{3}_{u+1}(x_1, x_2, x_3)$ of the Markov approximation cannot be substituted by the marginal $P'^{3}_{u+1}(x_1, x_2, x_3)$. This in turn prevents the common factor $P'^{3}_{u+1}(x_1, x_2, x_3)$ from being extracted in Equation A28 and summed over $\sum_{x_1,x_2,x_3}$ to unity. A different form of proof would therefore be required to establish Equation A18 for all n, as we conjecture.

Markov invariance: We prove that the Markov approximations of the distributions arising from intermixing and meiosis are identical:

$$\|Q_{u+1} - Q'_{u+1}\| = 0. \tag{A29}$$

To prove Equation A29, it is sufficient to prove that $P_{u+1}(x_i, x_{i+1}) = P'_{u+1}(x_i, x_{i+1})$ for all $i = 1, \ldots, n - 1$ since the Markov approximations Q_{u+1} and Q'_{u+1} are defined solely in terms of these joint distributions between adjacent sites.

We compute $P_{u+1}(x_i, x_{i+1})$ by marginalizing $P_{u+1}(x)$, as given in Equation A3:

$$P_{u+1}(x_i, x_{i+1}) = \sum_r G(r) \sum_{y^1,y^2|y_i^{S(r,i)}=x_i,\,y_{i+1}^{S(r,i+1)}=x_{i+1}} P_u(y_i^1, y_{i+1}^1)P_u(y_i^2, y_{i+1}^2).$$

We now split the sum over r into two. If $r_i = 0$ then there is no crossover between sites i and $i + 1$. In this case, $S(r, i) = S(r, i + 1)$, yielding that both sites in x originate from the same parent. If $r_i = 1$ then there is a crossover between sites i and $i + 1$. In this case, $S(r, i) \neq S(r, i + 1)$, yielding that each site in x originates from a different parent. Therefore

$$P_{u+1}(x_i, x_{i+1}) = \sum_{r|r_i=0} G(r)P_u(x_i, x_{i+1}) + \sum_{r|r_i=1} G(r)P_u(x_i)P_u(x_{i+1}).$$

Using the definition of $G(r)$ in Equation A2, it follows that $\sum_{r|r_i=0} G(r) = 1 - \theta_i$ and $\sum_{r|r_i=1} G(r) = \theta_i$. Consequently, $P_{u+1}(x_i, x_{i+1}) = (1 - \theta_i) \cdot P_u(x_i, x_{i+1}) + \theta_i \cdot P_u(x_i)P_u(x_{i+1})$. This result corresponds with the intuition that the offspring joint distribution over sites i and $i + 1$ is the average of the parent joint distribution and parent marginal distributions, weighted by the probability of a crossover and no crossover, respectively. By similar means, it can be shown that $P'_{u+1}(x_i, x_{i+1}) = (1 - \theta_i) \cdot P_u(x_i, x_{i+1}) + \theta_i \cdot P_u(x_i)P_u(x_{i+1})$, yielding the desired equality $P_{u+1}(x_i, x_{i+1}) = P'_{u+1}(x_i, x_{i+1})$. This proves Equation A29.

Properties of variation distance: The first property relates the variation distance between two multivariate distributions $A(x, y)$ and $B(x, y)$ to the variation distance between the two marginal distributions $A(x) = \sum_y A(x, y)$ and $B(x) = \sum_y B(x, y)$:

$$\begin{aligned}\|A(x, y) - B(x, y)\| &= \frac{1}{2}\sum_x\sum_y |A(x, y) - B(x, y)| \\ &\geq \frac{1}{2}\sum_x \left|\sum_y \{A(x, y) - B(x, y)\}\right| \\ &= \frac{1}{2}\sum_x |A(x) - B(x)| \\ &= \|A(x) - B(x)\|.\end{aligned}$$

The second property relates the variation distance between two mixture distributions $A(x) = \alpha A_1(x) + (1 - \alpha)A_2(x)$ and $B(x) = \alpha B_1(x) + (1 - \alpha)B_2(x)$ to the variation distances between the respective mixture elements:

$$\begin{aligned}\|A(x) - B(x)\| &= \frac{1}{2}\sum_x |A(x) - B(x)| \\ &= \frac{1}{2}\sum_x |\alpha(A_1(x) - B_1(x)) + (1 - \alpha)(A_2(x) - B_2(x))| \\ &\leq \frac{1}{2}\sum_x |\alpha(A_1(x) - B_1(x))| + \frac{1}{2}\sum_x |(1 - \alpha)(A_2(x) - B_2(x))| \\ &= \alpha\|A_1(x) - B_1(x)\| + (1 - \alpha)\|A_2(x) - B_2(x)\|.\end{aligned}$$