Modeling Haplotype Block Variation using Markov Chains

Greenspan G. and Geiger D.

Computer Science Department
Technion, Haifa 32000, Israel
Running title: Haplotype Blocks and Markov Chains

Key words: haplotype blocks, recombination hotspots, Markov models, linkage disequilibrium, LD mapping

Corresponding author: Gideon Greenspan, Computer Science Department, Technion, Technion City, Haifa 32000, Israel. Tel: +972 4 829 5537, Fax: +972 4 829 3900. Email: gdg@cs.technion.ac.il
Abstract

Models of background variation in genomic regions form the basis of linkage disequilibrium mapping methods. In this work we analyze a background model which 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 which assumes 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.
INTRODUCTION

Genetic mapping studies based on linkage disequilibrium (LD) require a model of the background variation in the region being examined. The study looks 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 $\chi^2$ 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 in the surrounding regions (Jeffreys et al. 2000; Templeton et al. 2000; Jeffreys et al. 2001; 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 (The 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 which 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 which are close together tend to exhibit a high degree of linkage disequilibrium that cannot be captured by an independent approximation.

In this paper, 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 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 paper 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 which are close together than those which 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; Greenspan and Geiger 2004b).

MATERIALS AND METHODS

Independent and Markov Approximations: Consider a genomic region which 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, labelled \( 1 \ldots r_j \). 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 as follows:

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

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 as follows:

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

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

**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 indices $j_1 \ldots j_n$ 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:

$$||A(z_1, \ldots, z_n) - B(z_1, \ldots, z_n)|| = \frac{1}{2} \sum_{z_1 \ldots z_n} |A(z_1, \ldots, z_n) - B(z_1, \ldots, z_n)|$$

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 in order to make one distribution equal to the other. For example, $||P - T||$ is the percentage of the population distributed as $P$ which 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

$$D = p_{11} - p_1 q_1 = p_1 q_2 - p_{12} = p_2 q_1 - p_{21} = p_{22} - p_2 q_2$$

(Devlin and Risch 1995). For example, if A and B are in perfect linkage equilibrium, then $p_{11} = p_1 q_1$, $p_{12} = p_1 q_2$, $p_{21} = p_2 q_1$ and $p_{22} = p_2 q_2$, and so $D = 0$. By comparison, the variation distance between $P$ and $T$ is:

$$||P - T|| = \frac{1}{2} (|p_{11} - p_1 q_1| + |p_{12} - p_1 q_2| + |p_{21} - p_2 q_1| + |p_{22} - p_2 q_2|)$$

$$= \frac{1}{2} (|D| + |D| + |D| + |D|) = 2|D|$$

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 the International Haplotype Mapping (HapMap) project to profile the error rates of the independent and Markov approximations for the human genome (The 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, (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 (Greenspan and Geiger 2004b; Zhang and Jin 2003). 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 4, 6, and 8 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 6 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 over 50% certainty.

**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 usefully be examined for the respective models, since $n$ markers spread over distance $d$ must be spaced at least $d/(n - 1)$ apart in order to be included in the averages $Y_{d,n}$ and $Z_{d,n}$. Table 2 compares the average figures over all 22 chromosomes for the six HaploBlock and HaploBlockFinder models. This table 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 $3 \leq n \leq 5$.

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 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 $n = 3, 4, 5$, respectively. To avoid a bias at short distances towards genomic regions with particularly high levels of variation, the graph in Figure 1 only shows 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 which 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 a 8% improvement compared to four 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 \( d = 350\text{kb} \) is equal to that at long distances, but is less accurate at distances between.

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 \( n = 3, 4, 5 \), 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 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 4, 6 and 8 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 approximations profiles for the three HaploBlockFinder partitions. Recall that the threshold specifies the percentage of the variation within each block which 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.

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 1 or 2 SNPs perform worse at short distances than at long distances, this relationship is reversed for groups of 4 SNPs or more. The values in Figure 6 are relative to baselines of 0.051, 0.123, 0.184 and 0.294 for 1, 2, 4 and 6 SNPs respectively.

The curves in Figures 4, 5 and 6 are labelled 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, 4, 5 and 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 are 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 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 UCSC 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 $0.506 \pm 0.281$. This contrasts with the independent approximation for the
HaploBlock model, with an average coefficient of $-0.705 \pm 0.292$.

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

The correlations coefficients for chromosomes 21 and 22 in Table 3 differ significantly from the mean values in many cases. This is because the HapMap data covers 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 to 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 anti-correlated 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 about 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 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 $\Delta Z/\Delta cM/Mb$ 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.

**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, (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 only affects 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 only represents 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.

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 multi-allelic 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, 4, 5 and 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 which crosses over from one parent to the other with probability $\theta_j$ between markers $j$ and $j+1$, where $\theta_j$ 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 which moves to a new parent with probability $\theta_j$ 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 only uses the first two parent haplotypes, producing the same offspring haplotype as meiosis. Differences only arise 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 \( \theta_j \) 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 \( Q_u(x_1, \ldots, x_n) = P_u(x_1) \prod_{i=1}^{n-1} P_u(x_{i+1}|x_i) \) 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 \( D_u(j) = 1 - \sum_{j} (P_u(x_j))^2 \) 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 \( n \leq 5 \):

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

(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 which 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:

21
\[ \|P_{u+1} - Q_{u+1}\| \leq \|P_{u+1} - P'_{u+1}\| + \|P'_{u+1} - Q'_u\| + \|Q'_u - Q_{u+1}\| \]

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 \(\|P'_{u+1} - Q'_u\|\), 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
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 only provides 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 the HapMap project, 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, in order to overcome this disadvantage. Nonetheless, since our empirical observations were based on averages over large numbers of sites, this shortcoming does not effect 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 a specific model of haplotype blocks. Nonetheless haplotype block based on statistical criteria offer other advantages over arbitrary groups of SNPs in terms of model simplicity and the selection of haplotype tagging SNPs (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 developing more precise criteria for determining when each plays a more dominant role. It is also desirable to ascertain whether a population must contain highly distinct haplotypes in order for the perturbation effect to be seen. On this point, recent research has found an abundance of common haplotypes which 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


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.


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: 858–75.


Minimal haplotype tagging. PNAS USA 100: 9900–5.

Templeton, A., A. Clark, K. Weiss, D. Nickerson, E. Boerwinkle et al., 2000
Recombinational and mutational hotspots within the human lipoprotein lipase gene.  

The International HapMap Consortium, 2003 The International HapMap Project.  

Tishkoff, S. A., and B. C. Verrelli, 2003 Role of evolutionary history on haplotype  
block structure in the human genome: implications for disease mapping.  

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.  

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.  

Wall, J. D., and J. K. Pritchard, 2003 Assessing the performance of the haplotype  

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  


APPENDIX

Here we prove in full the theoretical result outlined in the paper.

**Definitions:** Under meiotic recombination, each offspring haplotype over \( n \) sites is formed from two parent haplotypes \( y^1 = (y^1_1 \ldots y^1_n) \) and \( y^2 = (y^2_1 \ldots y^2_n) \). 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 \( r \):

\[
F(y^1, y^2, r) = y^{S(r,1)}_1 \ldots y^{S(r,n)}_n
\]

In Equation 2, \( S(r, i) \) 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 is an even number of recombinations up to site \( i \) then \( S(r, i) \) is 1, otherwise \( S(r, i) \) is 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 \( \theta_i \). We define the probability \( G(r) \) of a crossover vector \( r \) in terms of these pairwise probabilities:

\[
G(r_1, \ldots, r_{n-1}) = \prod_{i=1}^{n-1} \theta_i^{r_i} \cdot (1 - \theta_i)^{1-r_i}
\]

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 \( x \):

\[
P_{u+1}(x) = \sum_{y^1, y^2, r \mid F(y^1, y^2, r) = x} G(r) P_u(y^1) P_u(y^2)
\]

(4)

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 \( r \):

\[
F'(y^1, y^2, r) = y_{S'(r,1)}^1 \ldots y_{S'(r,n)}^n
\]

(5)

In Equation 5, \( S'(r, i) \) 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:

\[
P'_{u+1}(x) = \sum_{y^1 \ldots y^n, r \mid F'(y^1, \ldots, y^n, r) = x} G(r) \prod_{i=1}^{n} P_u(y^i)
\]

(6)

**Intermixing and Meiosis:** In this section, 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:
$$\left\| P_{u+1} - P'_{u+1} \right\| \leq \frac{1}{2} \left( \sum_{i=1}^{n-1} \theta_i \right)^2 \cdot \min \left( 1, \sum_{j=3}^{n} D_u(j) \right) \quad (7)$$

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 one or less crossovers, and let $R^+ = \{r \in R | \sum_j r_j \geq 2\}$ denote the subset representing two or more crossovers. Clearly, $R = R^- \cup R^+$ and $R^- \cap R^+ = \emptyset$. The frequency of haplotype $x$ after meiosis, given in Equation 4, can therefore be written as:

$$P_{u+1}(x) = \sum_{y^1, y^2, r \in R^- | F(y^1, y^2, r) = x} G(r) P_u(y^1) P_u(y^2) + \sum_{y^1, y^2, r \in R^+ | F(y^1, y^2, r) = x} G(r) P_u(y^1) P_u(y^2) \quad (8)$$

Similarly, the frequency of $x$ after intermixing, given in Equation 6, can be written as:

$$P'_{u+1}(x) = \sum_{y^1 \ldots y^n, r \in R^- | F'(y^1 \ldots y^n, r) = x} G(r) \prod_{i=1}^{n} P_u(y^i) + \sum_{y^1 \ldots y^n, r \in R^+ | F'(y^1 \ldots y^n, r) = x} G(r) \prod_{i=1}^{n} P_u(y^i) \quad (9)$$
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 9 as follows:

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

Since the sums in Equations 8 and 10 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:

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

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 11, we obtain:

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

We now derive the bound for \( ||P_{u+1} - P'_{u+1}|| \), as given by Equation 7. Let \([a= b]\) denote
the function that returns 1 if \(a = b\) and 0 otherwise, and define \([a \neq b] = 1 - [a = b]\).

Equation 12 is reformulated as follows:

\[
\left| P_{u+1} - P'_{u+1} \right| = \frac{1}{2} \sum_x \left| \sum_{y^1 \ldots y^n, x \in R^+} G(r) \prod_{i=1}^n P_u(y^i) \right. \\
\left. \cdot \left\{ [F(y^1, y^2, r) = x] - [F'(y^1, \ldots, y^n, r) = x] \right\} \right| \\
\leq \sum_{r \in R^+} G(r) \sum_{y^1 \ldots y^n} \prod_{i=1}^n P_u(y^i) \cdot \frac{1}{2} \sum_x \left| [F(y^1, y^2, r) = x] - [F'(y^1, \ldots, y^n, r) = x] \right| \\
= \sum_{r \in R^+} G(r) \sum_{y^1 \ldots y^n} \prod_{i=1}^n P_u(y^i) \cdot [F(y^1, y^2, r) \neq F'(y^1, \ldots, y^n, r)]
\tag{13}
\]

The last equality follows because if \(F(y^1, y^2, r) = F'(y^1, \ldots, y^n, r)\) then the expression \([F(y^1, y^2, r) = x] - [F'(y^1, \ldots, y^n, r) = x]\) is equal to 0 for all \(x\), and if \(F(y^1, y^2, r) \neq F'(y^1, \ldots, y^n, r)\), then \([F(y^1, y^2, r) = x] - [F'(y^1, \ldots, y^n, r) = x]\) is equal to 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 \([F(y^1, y^2, r) \neq F'(y^1, \ldots, y^n, r)]\) is 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 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:

\[
[F(y^1, y^2, r) \neq F'(y^1, \ldots, y^n, r)] \leq \sum_{j=3}^n [y_j^{S(r,j)} \neq y_j^{S'(r,j)}]
\tag{14}
\]
Equations 13 and 14 yield:

\[ \left| \left| P_{u+1} - P'_{u+1} \right| \right| \leq \sum_{r \in R^+} G(r) \sum_{j=3}^{n} \sum_{y^1 \ldots y^n} \prod_{i=1}^{n} P_u(y^i) \cdot [y_j^{S(r,j)} \neq y_j^{S'(r,j)}] \tag{15} \]

Since, in the worst case, every site from the third onwards 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:

\[ \left| \left| P_{u+1} - P'_{u+1} \right| \right| \leq \sum_{r \in R^+} G(r) \sum_{j=3}^{n} D_u(j) \tag{16} \]

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 \( \left| \left| P_{u+1} - P'_{u+1} \right| \right| \) from Equation 13:

\[ \left| \left| P_{u+1} - P'_{u+1} \right| \right| \leq \sum_{r \in R^+} G(r) \sum_{y^1 \ldots y^n} \prod_{i=1}^{n} P_u(y^i) = \sum_{r \in R^+} G(r) \tag{17} \]

Finally, using the probability \( G(r) \) of a crossover vector \( r \) (Equation 3), we bound \( \sum_{r \in R^+} G(r) \) by summing the probability of every possible pair of crossovers.
Equations 16, 17 and 18 yield the bound for $||P_{u+1} - P'_{u+1}||$, given by Equation 7.

**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 $P'_{u+1}(x)$. In this section we prove that for $n \leq 5$:

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

(19)

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 6 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 $r$:

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

(20)

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

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

$U(r, k) = \max \{i|S'(r, i) = k\}$
In Equation 20, the functions $L(r, k)$ and $U(r, k)$ denote respectively the first and last sites in the offspring haplotype which 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 $P_u(x_{L(r,k)}, \ldots, x_{U(r,k)}) = \sum x_{1\ldots L(r,k)-1} x_{U(r,k)+1\ldots l} P_u(x_1, \ldots, x_l)$.

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 only allowed over sites 1 to $i$. In other words, $P'^i_{u+1}$ is the result of intermixing on $P_u$ if all values $\theta_1 \ldots \theta_{n-1}$ 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 $n$ are 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 only occur between sites 1 to $i$, i.e. $R^i = \{ r \in R | r_i = 0 \ldots r_{n-1} = 0 \}$. 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 $P_{u+1}^i(x)$ of haplotype $x$ after intermixing over sites $1 \ldots i$ is analogous to $P_{u+1}^i(x)$, given in Equation 20:

\[
P_{u+1}^i(x_1, \ldots, x_n) = \sum_{r \in R^i} G^i(r) \prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)})
\]

\[
= \sum_{r \in R^i} G^i(r) \left( \prod_{k=1}^{S'(r,n)-1} P_u(x_{(r,k)}) \right) P_u(x_{L(r,S'(r,n))}, \ldots, x_n)
\]  

(21)

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

\[
P_{u+1}^{i+1}(x) = \sum_{r \in R^{i+1} | r_i = 0} G^{i+1}(r) \prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)}) + \sum_{r \in R^{i+1} | r_i = 1} G^{i+1}(r) \prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)})
\]

\[
= (1 - \theta_i) \sum_{r \in R^{i+1} | r_i = 0} G^i(r) \prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)})
\]

\[
+ \theta_i \sum_{r \in R^{i+1} | r_i = 1} G^i(r) \left( \prod_{k=1}^{S'(r,n)-1} P_u(x_{(r,k)}) \right) P_u(x_{L(r,S'(r,n))}, \ldots, x_n)
\]

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,

\[
P_{u+1}^{i+1}(x) = (1 - \theta_i) \sum_{r \in R^i} G^i(r) \prod_{k=1}^{S'(r,n)} P_u(x_{(r,k)})
\]
\[
\sum_{r \in R^{i+1} | r_i = 1} G^i(r) \left( \prod_{k=1}^{S'(r,n)-1} P_u(x_{(r,k)}) \right) P_u(x_{L(r,S'(r,n))}, \ldots, x_n)
\]

\[
(1 - \theta_i) P_{u+1}^i (x_1, \ldots, x_n)
\]  

\[
\sum_{r \in R^i} G^i(r) \left( \prod_{k=1}^{S'(r,n)-2} P_u(x_{(r,k)}) \right) P_u(x_{L(r,S'(r,n)-1)}, \ldots, x_i) P_u(x_{i+1}, \ldots, x_n)
\]

We now replace the sum over \( r \in R^{i+1} | r_i = 1 \) 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 \( i + 1 \):

\[
P_{u+1}^{i+1}(x) = (1 - \theta_i) P_{u+1}^i (x_1, \ldots, x_n)
\]

\[
\sum_{r' \in R^i} G^i(r') \left( \prod_{k=1}^{S'(r',n)-1} P_u(x_{(r',k)}) \right) P_u(x_{L(r',S'(r',n))}, \ldots, x_i) P_u(x_{i+1}, \ldots, x_n)
\]

\[
= (1 - \theta_i) \cdot P_{u+1}^i (x_1, \ldots, x_n) + \theta_i \cdot P_{u+1}^i (x_1, \ldots, x_i) P_u(x_{i+1}, \ldots, x_n)
\]

We have replaced \( G^i(r) \) with \( G^i(r') \) in the transformation from Equation 22 to Equation 23 since the function \( G^i \) is not affected by crossovers after site \( i \). The function \( S'(r,n) \) in Equation 22 counts the total number of crossovers represented by vector \( r \). It is replaced by \( S'(r',n) + 1 \) in Equation 23 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 22 is replaced by the product \( \prod_{k=1}^{S'(r',n)-1} P_u(x_{(r',k)}) \) in Equation 23 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 22 is replaced with \( L(r', S'(r', n)) \) in Equation 23 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 $P'_{i+1}$ is the result of intermixing only up to site $i$, so its marginal $P'_i(x_{i+1}, \ldots, x_n)$ over sites $i+1 \ldots n$ is the same as the parent marginal $P_u(x_{i+1}, \ldots, x_n)$. Consequently, Equation 23 implies:

$$P'_{i+1}(x) = (1 - \theta_i) \cdot P'_{i+1}(x_1, \ldots, x_n) + \theta_i \cdot P'_{i+1}(x_1, \ldots, x_i)P'_{i+1}(x_{i+1}, \ldots, x_n) \quad (24)$$

Equation 24 states that the effect of introducing the additional crossover point between sites $i$ and $i + 1$ is to reconstitute a proportion $\theta_i$ of the population from the marginal distributions on either side of the crossover point, leaving the remaining $1 - \theta_i$ proportion untouched. Equation 24 also holds in the following marginal form by summing over $x_1, \ldots, x_{i-1}, x_{i+2}, \ldots, x_n$:

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

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

$$Q'_{i+1}(x_1, \ldots, x_n) = P'_{i+1}(x_1) \prod_{j=1}^{n-1} P'_{i+1}(x_{j+1}|x_j) \quad (25)$$

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

Equation 26 states the analogous result for the series of Markov approximations $Q^{n+1}$ does not affect the marginal allele frequencies for any individual site. Similarly, we replaced as Equation 24 states for the series of distributions $P^{n}$ establishes this inequality, we prove that for $1 \leq i \leq n$.

Recall that we aim to prove $(x_{j} \mid x_{j}) = (1 - \theta_{i})P^{n+1}_{u+1}(x_{i}, x_{i+1}) + \theta_{i}P^{n+1}_{u+1}(x_{i+1})P^{n}_{u+1}(x_{i+1}) + \prod_{j=i}^{n-1}P^{n}_{u+1}(x_{j+1} \mid x_{j})$

$Q^{n+1}_{u+1}(x) = (1 - \theta_{i})Q^{n}_{u+1}(x_{1}, \ldots, x_{n}) + \theta_{i}Q^{n}_{u+1}(x_{1}, \ldots, x_{i}) \cdot Q^{n}_{u+1}(x_{i+1}, \ldots, x_{n})$  \hspace{1cm}  \text{(26)}$Q^{n+1}_{u+1}(x)$

We replaced $P^{n+1}_{u+1}(x_{i})$ with $P^{n}_{u+1}(x_{i})$ at several points above since the intermixing process does not affect the marginal allele frequencies for any individual site. Similarly, we replaced $P^{n+1}_{u+1}(x_{j+1} \mid x_{j})$ with $P^{n}_{u+1}(x_{j+1} \mid x_{j})$ for any $j \neq i$ since the additional crossover permitted between sites $i$ and $i + 1$ only affects marginal distributions containing both $x_{i}$ and $x_{i+1}$. Equation 26 states the analogous result for the series of Markov approximations $Q^{n+1}_{u+1} \ldots Q^{n}_{u+1}$ as Equation 24 states for the series of distributions $P^{n+1}_{u+1} \ldots P^{n}_{u+1}$.

Recall that we aim to prove $\|P_{u+1}^{n} - Q_{u+1}^{n+1}\| \leq \|P_{u} - Q_{u}\|$ for $n \leq 5$. Since $P_{u+1}^{n} = P_{u}$ and $P_{u+1}^{n} = P_{u+1}^{n}$, this inequality can be expressed as $\|P_{u+1}^{n} - Q_{u+1}^{n+1}\| \leq \|P_{u+1}^{n} - Q_{u+1}^{n}\|$. To establish this inequality, we prove that for $1 \leq i \leq n - 1$: 

$|Q_{u+1}^{n+1}(x) - P_{u+1}^{n+1}(x)| = |P_{u+1}^{n}(x) - P_{u+1}^{n+1}(x)|$
\[
\left\| P_{u+1}^{i+1} - Q_{u+1}^{i+1} \right\| \leq \left\| P_{u+1}^{i} - Q_{u+1}^{i} \right\| \quad (27)
\]

We split the proof of Equation 27 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 \( \left\| A(x, y) - B(x, y) \right\| \geq \left\| A(x) - B(x) \right\| \). 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 \( \left\| A(x) - B(x) \right\| \leq \alpha \left\| A_1(x) - B_1(x) \right\| + (1 - \alpha) \left\| A_2(x) - B_2(x) \right\| \). Proofs of these two properties are provided at the end of the Appendix.

For \( i = 1 \), we prove Equation 27 by rewriting \( P_{u+1}^{2} \) and \( Q_{u+1}^{2} \) in terms of \( P_{u+1}^{1} \) and \( Q_{u+1}^{1} \), using the recurrence relations in Equations 24 and 26:

\[
\begin{align*}
P_{u+1}^{2} (x) &= (1 - \theta_1) \cdot P_{u+1}^{1} (x_1, \ldots, x_n) + \theta_1 \cdot P_{u+1}^{1} (x_1) \cdot P_{u+1}^{1} (x_2, \ldots, x_n) \\
Q_{u+1}^{2} (x) &= (1 - \theta_1) \cdot Q_{u+1}^{1} (x_1, \ldots, x_n) + \theta_1 \cdot Q_{u+1}^{1} (x_1) \cdot Q_{u+1}^{1} (x_2, \ldots, x_n) \\
&\quad = (1 - \theta_1) \cdot Q_{u+1}^{1} (x_1, \ldots, x_n) + \theta_1 \cdot P_{u+1}^{1} (x_1) \cdot Q_{u+1}^{1} (x_2, \ldots, x_n)
\end{align*}
\]

The last equality follows because the marginal distribution for an individual site is identical for both \( P_{u+1}^{1} \) and its Markov approximation \( Q_{u+1}^{1} \). The proof of Equation 27 for \( i = 1 \)
is completed using the two properties of variation distance:

\[
\left| P_{u+1}^2 - Q_{u+1}^2 \right| \leq (1 - \theta_1) \cdot \left| P_{u+1}^1 - Q_{u+1}^1 \right| +
\theta_1 \cdot \frac{1}{2} \sum_{x_1 \ldots x_n} \left| P_{u+1}^1(x_1)P_{u+1}^1(x_2, \ldots, x_n) - P_{u+1}^1(x_1)Q_{u+1}^1(x_2, \ldots, x_n) \right|
\]

\[
= (1 - \theta_1) \cdot \left| P_{u+1}^1 - Q_{u+1}^1 \right| +
\theta_1 \cdot \frac{1}{2} \sum_{x_1} P_{u+1}^1(x_1) \sum_{x_2 \ldots x_n} \left| P_{u+1}^1(x_2, \ldots, x_n) - Q_{u+1}^1(x_2, \ldots, x_n) \right|
\]

\[
\leq (1 - \theta_1) \cdot \left| P_{u+1}^1 - Q_{u+1}^1 \right| + \theta_1 \cdot \left| P_{u+1}^1 - Q_{u+1}^1 \right|
\]

\[
= \left| P_{u+1}^1 - Q_{u+1}^1 \right|
\]

For \(i = 2\), the proof of Equation 27 proceeds similarly:

\[
P_{u+1}^3(x) = (1 - \theta_2) \cdot P_{u+1}^2(x_1, \ldots, x_n) + \theta_2 \cdot P_{u+1}^2(x_1, x_2) \cdot P_{u+1}^2(x_3, \ldots, x_n)
\]

\[
Q_{u+1}^3(x) = (1 - \theta_2) \cdot Q_{u+1}^2(x_1, \ldots, x_n) + \theta_2 \cdot Q_{u+1}^2(x_1, x_2) \cdot Q_{u+1}^2(x_3, \ldots, x_n)
\]

\[
= (1 - \theta_2) \cdot Q_{u+1}^2(x_1, \ldots, x_n) + \theta_2 \cdot P_{u+1}^2(x_1, x_2) \cdot Q_{u+1}^2(x_3, \ldots, x_n)
\] (28)

The last equality follows since the joint distribution over any two adjacent sites is unchanged by the Markov approximation. The proof of Equation 27 for \(i = 2\) is completed using the two properties of variation distance:

\[
\left| P_{u+1}^3 - Q_{u+1}^3 \right| \leq (1 - \theta_2) \cdot \left| P_{u+1}^2 - Q_{u+1}^2 \right| +
\theta_2 \cdot \left| P_{u+1}^2 - Q_{u+1}^2 \right|
\]
\[ \theta_2 \cdot \frac{1}{2} \sum_{x_1, \ldots, x_n} |P_{u+1}^2(x_1, x_2)P_{u+1}^2(x_3, \ldots, x_n) - P_{u+1}^2(x_1, x_2)Q_{u+1}^2(x_3, \ldots, x_n)| \]

\[ = (1 - \theta_2) \cdot \left| P_{u+1}^2 - Q_{u+1}^2 \right| + \theta_2 \cdot \left| P_{u+1}^2 - Q_{u+1}^2 \right| \]

\[ \leq (1 - \theta_2) \cdot \left| P_{u+1}^2 - Q_{u+1}^2 \right| + \theta_2 \cdot \left| P_{u+1}^2 - Q_{u+1}^2 \right| \]

\[ = \left| P_{u+1}^2 - Q_{u+1}^2 \right| \]

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 \( \left| P_{u+1}^{i+1} - Q_{u+1}^{i+1} \right| \leq \left| P_{u+1}^i - Q_{u+1}^i \right| \) and therefore that \( \left| P_{u+1}^i - Q_{u+1}^i \right| \leq \left| P_u - Q_u \right| \) for \( n \leq 5 \), as stated in Equation 19.

For \( n > 5 \), this method breaks down in Equation 28 for \( i = 3 \) since the marginal distribution \( Q_{u+1}^3(x_1, x_2, x_3) \) of the Markov approximation cannot be substituted by the marginal \( P_{u+1}^3(x_1, x_2, x_3) \). This in turn prevents the common factor \( P_{u+1}^3(x_1, x_2, x_3) \) from being extracted in Equation 29 and summed over \( \sum x_1, x_2, x_3 \) to unity. A different form of proof would therefore be required to establish Equation 19 for all \( n \), as we conjecture.

**Markov Invariance:** In this section we prove that the Markov approximations of the distributions arising from intermixing and meiosis are identical:

\[ \left| Q_{u+1} - Q'_{u+1} \right| = 0 \quad (30) \]

To prove Equation 30, 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 4:

\[
P_{u+1}(x_i, x_{i+1}) = \sum_r G(r) \sum_{y^1, y^2: y_i^{S(r, i)} = x_i, y_{i+1}^{S(r, i+1)} = x_{i+1}} P_u(y^1_i, y^1_{i+1}) P_u(y^2_i, y^2_{i+1})
\]

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 3, 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 30.

**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 \( B(x) = \sum_y B(x, y) \):

\[
||A(x, y) - B(x, y)|| = \frac{1}{2} \sum_x \sum_y |A(x, y) - B(x, y)| \\
\geq \frac{1}{2} \sum_x \left| \sum_y \{A(x, y) - B(x, y)\} \right| \\
= \frac{1}{2} \sum_x |A(x) - B(x)| \\
= ||A(x) - B(x)||
\]

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:

\[
||A(x) - B(x)|| = \frac{1}{2} \sum_x |A(x) - B(x)| \\
= \frac{1}{2} \sum_x |\alpha (A_1(x) - B_1(x)) + (1 - \alpha) (A_2(x) - B_2(x))| \\
\leq \frac{1}{2} \sum_x |\alpha (A_1(x) - B_1(x))| + \frac{1}{2} \sum_x |(1 - \alpha) (A_2(x) - B_2(x))| \\
= \alpha ||A_1(x) - B_1(x)|| + (1 - \alpha) ||A_2(x) - B_2(x)||
\]
Table 1: Summary for each chromosome of SNPs and the HaploBlock model with up to four variants. All distances are in kb.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Length</th>
<th>Count</th>
<th>Mean spacing</th>
<th>Count</th>
<th>Mean spacing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>245416</td>
<td>47618</td>
<td>5.15</td>
<td>4587</td>
<td>53.47</td>
</tr>
<tr>
<td>2</td>
<td>243363</td>
<td>68659</td>
<td>3.54</td>
<td>5604</td>
<td>43.42</td>
</tr>
<tr>
<td>3</td>
<td>199162</td>
<td>43880</td>
<td>4.54</td>
<td>4132</td>
<td>48.21</td>
</tr>
<tr>
<td>4</td>
<td>191628</td>
<td>42213</td>
<td>4.54</td>
<td>4059</td>
<td>47.19</td>
</tr>
<tr>
<td>5</td>
<td>180747</td>
<td>47651</td>
<td>3.79</td>
<td>4070</td>
<td>44.41</td>
</tr>
<tr>
<td>6</td>
<td>170674</td>
<td>37013</td>
<td>4.61</td>
<td>3567</td>
<td>47.84</td>
</tr>
<tr>
<td>7</td>
<td>158508</td>
<td>35460</td>
<td>4.47</td>
<td>3354</td>
<td>47.24</td>
</tr>
<tr>
<td>8</td>
<td>146201</td>
<td>44926</td>
<td>3.25</td>
<td>3536</td>
<td>41.34</td>
</tr>
<tr>
<td>9</td>
<td>136199</td>
<td>30902</td>
<td>4.41</td>
<td>2812</td>
<td>48.42</td>
</tr>
<tr>
<td>10</td>
<td>134982</td>
<td>29391</td>
<td>4.59</td>
<td>2972</td>
<td>45.38</td>
</tr>
<tr>
<td>11</td>
<td>134292</td>
<td>37281</td>
<td>3.60</td>
<td>3106</td>
<td>43.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------</td>
<td>-------</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>12</td>
<td>131969</td>
<td>34894</td>
<td>3.78</td>
<td>3093</td>
<td>42.66</td>
</tr>
<tr>
<td>13</td>
<td>96204</td>
<td>24138</td>
<td>3.99</td>
<td>2201</td>
<td>43.68</td>
</tr>
<tr>
<td>14</td>
<td>87070</td>
<td>24104</td>
<td>3.61</td>
<td>2098</td>
<td>41.51</td>
</tr>
<tr>
<td>15</td>
<td>81870</td>
<td>22762</td>
<td>3.60</td>
<td>1966</td>
<td>41.61</td>
</tr>
<tr>
<td>16</td>
<td>90025</td>
<td>21516</td>
<td>4.18</td>
<td>2010</td>
<td>44.76</td>
</tr>
<tr>
<td>17</td>
<td>81652</td>
<td>21511</td>
<td>3.80</td>
<td>1950</td>
<td>41.89</td>
</tr>
<tr>
<td>18</td>
<td>76096</td>
<td>20545</td>
<td>3.70</td>
<td>1866</td>
<td>40.68</td>
</tr>
<tr>
<td>19</td>
<td>63742</td>
<td>15265</td>
<td>4.18</td>
<td>1564</td>
<td>40.69</td>
</tr>
<tr>
<td>20</td>
<td>63623</td>
<td>10794</td>
<td>5.89</td>
<td>1402</td>
<td>45.36</td>
</tr>
<tr>
<td>21</td>
<td>36965</td>
<td>17071</td>
<td>2.17</td>
<td>1279</td>
<td>27.62</td>
</tr>
<tr>
<td>22</td>
<td>34877</td>
<td>15520</td>
<td>2.25</td>
<td>1313</td>
<td>26.41</td>
</tr>
<tr>
<td>---</td>
<td>-------</td>
<td>-------</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Overall</td>
<td>2785266</td>
<td>693114</td>
<td>4.02</td>
<td>62541</td>
<td>44.49</td>
</tr>
</tbody>
</table>
Table 2: Summary over all chromosomes of different haplotype block models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Block count</th>
<th>Mean spacing (kb)</th>
<th>Mean variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>HaploBlock max 4 variants</td>
<td>62541</td>
<td>44.49</td>
<td>3.60</td>
</tr>
<tr>
<td>HaploBlock max 6 variants</td>
<td>48749</td>
<td>57.07</td>
<td>4.84</td>
</tr>
<tr>
<td>HaploBlock max 8 variants</td>
<td>43364</td>
<td>64.16</td>
<td>5.68</td>
</tr>
<tr>
<td>HaploBlockFinder 90% coverage</td>
<td>88038</td>
<td>31.61</td>
<td>5.75</td>
</tr>
<tr>
<td>HaploBlockFinder 80% coverage</td>
<td>53207</td>
<td>52.34</td>
<td>8.79</td>
</tr>
<tr>
<td>HaploBlockFinder 70% coverage</td>
<td>39575</td>
<td>70.31</td>
<td>11.51</td>
</tr>
</tbody>
</table>
Table 3: Correlation between recombination rates and error measures for sliding window over individual chromosomes. The haplotype block column refers to the HaplotypeBlock model with up to four variants.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Haplotype blocks</th>
<th>Individual SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Markov</td>
<td>Independent</td>
</tr>
<tr>
<td>1</td>
<td>0.519</td>
<td>-0.794</td>
</tr>
<tr>
<td>2</td>
<td>0.531</td>
<td>-0.791</td>
</tr>
<tr>
<td>3</td>
<td>0.618</td>
<td>-0.797</td>
</tr>
<tr>
<td>4</td>
<td>0.665</td>
<td>-0.788</td>
</tr>
<tr>
<td>5</td>
<td>0.474</td>
<td>-0.851</td>
</tr>
<tr>
<td>6</td>
<td>0.611</td>
<td>-0.766</td>
</tr>
<tr>
<td>7</td>
<td>0.132</td>
<td>-0.887</td>
</tr>
<tr>
<td>8</td>
<td>0.803</td>
<td>-0.690</td>
</tr>
<tr>
<td>9</td>
<td>-0.008</td>
<td>-0.675</td>
</tr>
<tr>
<td>10</td>
<td>0.522</td>
<td>-0.823</td>
</tr>
<tr>
<td>11</td>
<td>0.468</td>
<td>-0.762</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>12</td>
<td>0.826</td>
<td>-0.806</td>
</tr>
<tr>
<td>13</td>
<td>0.816</td>
<td>-0.949</td>
</tr>
<tr>
<td>14</td>
<td>0.455</td>
<td>-0.909</td>
</tr>
<tr>
<td>15</td>
<td>0.833</td>
<td>-0.848</td>
</tr>
<tr>
<td>16</td>
<td>0.522</td>
<td>-0.734</td>
</tr>
<tr>
<td>17</td>
<td>0.241</td>
<td>-0.812</td>
</tr>
<tr>
<td>18</td>
<td>0.571</td>
<td>-0.927</td>
</tr>
<tr>
<td>19</td>
<td>0.595</td>
<td>-0.619</td>
</tr>
<tr>
<td>20</td>
<td>0.941</td>
<td>-0.503</td>
</tr>
<tr>
<td>21</td>
<td>0.060</td>
<td>0.355</td>
</tr>
<tr>
<td>22</td>
<td>-0.069</td>
<td>-0.142</td>
</tr>
</tbody>
</table>

1-22 Mean  0.506  -0.705  -0.435  -0.666  
1-22 S.D.  0.281  0.292  0.423  0.358  

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20 Mean</td>
<td>0.557</td>
<td>-0.787</td>
<td>-0.477</td>
<td>-0.743</td>
</tr>
<tr>
<td>20 S.D.</td>
<td>0.238</td>
<td>0.107</td>
<td>0.355</td>
<td>0.207</td>
</tr>
</tbody>
</table>
Figure 1: Distance profile of Markov approximation for HaploBlock blocks
Figure 2: Distance profile of independent approximation for HaploBlock blocks
Figure 3: Comparison of distance profiles for HaploBlock blocks and individual SNPs
Figure 4: Comparison of Markov distance profiles for HaploBlock models
Figure 5: Comparison of Markov distance profiles for HaploBlockFinder models
Figure 6: Comparison of Markov distance profiles for SNP groupings
Figure 7: Position profiles for individual SNP models over chromosome 11
Figure 8: Position profiles for haplotype block models over chromosome 11
Figure 9: Effect of heterozygosity on average best-fit gradient between Markov error and recombination rates.
Figure 10: Meiotic recombination
Figure 11: Intermixing