Genomic Haplotype Blocks May Not Accurately Reflect Spatial Variation in Historic Recombination Intensity

Koen J. F. Verhoeven* and Katy L. Simonsen†

*Computational Genomics and Department of Agronomy, Purdue University; †Department of Statistics, Purdue University

Recently, genomic data have revealed a “block-like” structure of haplotype diversity on human chromosomes. This structure is anticipated to facilitate gene mapping studies, because strong associations among loci within a block may allow haplotype variation to be tagged with a limited number of markers. But its usefulness to mapping efforts depends on the consistency of the block structure within and among populations, which in turn depends on how the block structure arises. Recombination hot spots are generally thought to underlie the block structure, but haplotype blocks can also develop stochastically under random recombination, in which case the block structure will show limited consistency among populations. Using coalescent models, which we up-scaled to simulate the evolution of haplotypes with many markers at fixed distances, we show that the relationship between block boundaries and historic recombination intensity may be surprisingly weak. The majority of historic recombinations do not leave a footprint in present-day linkage disequilibrium patterns, and the block structure is sensitive to factors that affect the timing of recombination relative to marker mutation events in the genealogy, such as marker frequency bias and historic population size changes. Our results give insight into the potential of stochastic events to affect haplotype block structure, which can limit the usefulness of the block structure to mapping studies.

Introduction

An increasing amount of genomic data collected at the population level reveals that human chromosomes can be parsed into segments that show limited haplotype diversity, termed haplotype blocks (Daly et al. 2001; Patil et al. 2001; Dawson et al. 2002; Gabriel et al. 2002; Phillips et al. 2003). Within such blocks, loci show high levels of linkage disequilibrium (LD), that is, the alleles at different loci are not randomly assorted but allelic states are highly correlated. Blocks are separated by areas of low LD. There is currently great interest in the practical implications of this LD block structure for gene-mapping studies. Following the haplotype block model, sequence variation within extended blocks may be tagged by a few well-chosen SNPs (single nucleotide polymorphisms), allowing for efficient tests of haplotype-trait associations and significantly reducing the genotyping effort necessary for genome-wide mapping studies (Johnson et al. 2001). Driven by this aim of facilitating gene mapping, the human genome is screened for common patterns of sequence variation to produce a genomic map of high-LD blocks (the International HapMap Project; Int HapMap Consortium 2003).

Because recombination is the key factor that gradually disrupts allelic associations after an LD-generating event (such as a population bottleneck or population admixture), clustering of recombination in small hot spots could define the boundaries of larger haplotype blocks. Recombination hot spots have been modeled explicitly (Goldstein 2001; Nachman 2002; Jeffreys et al. 2004). This intuitively attractive hypothesis is supported by different lines of evidence, including the demonstration of spatial variation in recombination rates by comparing genetic and physical maps (Yu et al. 2001); coincidence of fine-scale recombination hot spots with haplotype block boundaries in sperm-typing studies (Jeffreys, Kauppi, and Neumann 2001; May et al. 2002; Kauppi, Sajantila, and Jeffreys 2003); and better fit with empirically observed LD patterns of models that incorporate recombination clustering compared to uniform recombination models in simulation studies (Reich et al. 2002; Wall and Pritchard 2003a).

However, simulation studies also show that the LD block structure can be affected by a range of factors other than recombination rate variability, such as marker density (Wang et al. 2002; Phillips et al. 2003; Wall and Pritchard 2003a), population demographic history (Wang et al. 2002; Stumpf and Goldstein 2003; Zhang et al. 2003; Anderson and Slatkin 2004), gene conversion (Przeworski and Wall 2001), ascertainment bias in marker selection (Akely et al. 2003; Phillips et al. 2003; Zhang et al. 2003), and, conceivably, by stochastic clustering of random recombinations (Subrahmanyan et al. 2001). And although modeling recombination hot spots does result in corresponding LD block boundaries, blocks also arise in areas of low recombination (Wall and Pritchard 2003b; Zhang et al. 2003). From these studies it is evident that a block-like LD structure can develop in the absence of recombination rate heterogeneity. The recombination hot spot hypothesis implies a strong relationship between recombination localization and haplotype block boundaries, but the observation that block boundaries arise readily under uniform recombination challenges the assumption that hot spots are responsible for most of the observed block-like LD patterns. This finding is relevant for the haplotype-tagging approach to gene mapping. The approach is most useful when the block structure is consistent within and between populations, which can be expected if the structure is biologically determined (via recombination hot spots), but is not obvious when the structure is mainly determined stochastically (Wang et al. 2002; Phillips et al. 2003), in which case the structure is shared only to the extent that the populations have a common history.

Using a new coalescent-based program that allows modeling of the genealogy of haplotypes consisting of
many markers at fixed distances, we simulated the evolution of haplotypes under a uniform recombination rate in order to investigate the relationship between historic recombination and present-day haplotype blocks. We show that the relation between historic recombination frequency and block boundaries can be quite weak, which emphasizes the potential of other stochastic events to shape LD blocks. We stress the impact on block structure of the timing of mutation relative to recombination events, and we illustrate this impact via the effects of historic population bottlenecks and of marker frequency constraints.

Materials and Methods

Coalescent Models

Under the assumption of neutrality, the coalescent process models the history of a population sample of sequences (Hudson 1990). It randomly generates a possible genealogy (ancestral recombination graph) for the sequences. Starting at the present day the process proceeds backward in time, allowing a succession of coalescent and recombination events to shape the genealogy. The mean waiting times (expressed typically in units of population size) between successive events and their relative probabilities are functions of the number of lineages present and the recombination rates between loci. Coalescent events join two randomly chosen lineages to form a common ancestor. Recombination events split a lineage in two, with two segments of a randomly chosen individual having different ancestors on each side of the recombination position; the main consequence of recombination is therefore that adjacent loci can have different genealogies. The process terminates when only one sequence, the most recent common ancestor of the entire sample, remains. After a genealogy is generated, mutations are superimposed on the branches according to a mutation model that is appropriate to the type of data (such as SNP haplotypes, full DNA sequences, or microsatellites).

Coalescent Program

We simulated genealogies based on a coalescent model with recombination (Kaplan and Hudson 1985; Simonsen and Churchill 1997). This process has been made more efficient to permit the simulation of thousands of loci by exploiting the sparsity of the Markov Chain structure (Simonsen et al. in preparation). Our implementation differs from that of Hudson (1983) in that recombination rates between each pair of adjacent loci are specified, and multiple recombination events are permitted at any given position. This permits a more realistic simulation of a genome with markers at pre-specified locations. For every genealogy, we recorded the actual number of recombination events that occurred at each position.

Population bottlenecks were imposed on the genealogies via modification of branch lengths in standard ways (Griffiths and Tavaré 1994). Briefly, if \( U \) is a coalescence time generated by a constant population size model, then the time \( T \) under a bottleneck model is

\[
T = \begin{cases} 
U & 0 < U \leq T_B \\
BU + TB(1-B) & T_B < U \leq T_B + \frac{TB}{B} \\
U - LB \left( \frac{1}{B} - 1 \right) & U > T_B + \frac{TB}{B} 
\end{cases}
\]

where \( TB \) is the amount of time since the bottleneck ended, \( LB \) is the duration of the bottleneck, and \( B \) is the fraction to which the size of the population is reduced.

Biallelic markers were simulated by imposing exactly one mutation at each locus, on a branch chosen with probability proportional to its length, subject to specified constraints on marker frequencies. For example, if minimum allele frequencies were specified as greater than or equal to 0.2, then branches leading to fewer than 0.2\( n \) or more than 0.8\( n \) sequences were not eligible for mutation.

The coalescent simulation program is available at http://www.stat.purdue.edu/~simonsen/Argsos/.

Simulation Settings

Unless stated otherwise, we simulated genealogies for 1-Mb segments with equidistant markers at 1-kb (1,000 markers) or 5-kb (200 markers) intervals, assuming a uniform recombination rate of 1 cM/Mb (Kruglyak 1999). In all simulations the basic effective population size was 10,000. To avoid unwarranted stochastic variation among models, the effects of marker frequency and population demographic history on haplotype blocks were assessed using the same underlying genealogies (20 iterations) for all models. Based on the constant population size model, the genealogies were compressed over part of the time range to fit the bottleneck scenarios and were subsequently subjected to the various mutation models that constrained the minimum allele frequencies. During a bottleneck, population size decreased to an effective population size of 10 for 50 generations. Note that these severe bottleneck models are not intended to reflect human demographic history, but to illustrate the effect on haplotype blocks of the relative timing of mutation and recombination events.

Linkage Disequilibrium and Haplotype Blocks

We used Lewontin’s \( D' \) (Lewontin 1964) as a measure of pairwise LD, and we defined a haplotype block as a contiguous region of at least three loci in which \( |D'| \geq 0.9 \) for all pairs of loci within the block (conform Phillips et al. 2003). In the simulated example with low sample size we also report Hill and Robertson’s \( r^2 \) index of LD (Hill and Robertson 1968), which is less sensitive to missing genotypes. To avoid overlapping blocks, we used a greedy algorithm (Zhang and Jin 2003) to maximize block length. With loci at equal distances of 1 kb, block length (in kb) was taken to be equal to the number of loci contained in the block.

Results and Discussion

The relation between recombinations and observed haplotype blocks can be weakened by genetic drift, removing recombinant genotypes from the population due to chance (Zhang et al. 2003). But even when only those recombination events are considered that involve the ancestors of a present-day population sample, block boundaries can be poorly...
lead to block boundaries. For instance, when the recombination rate is high, block boundaries can develop in the presence of mutation events, block boundaries are not affected by subsequent recombinations that can revert back to previous states at some loci. Recombination also affects the impact on present-day LD of a recombination event is related to local recombination frequency (fig. 1). Using a minimum-\(D\) \(2\) block definition, and without recurrent mutations, block boundaries can only develop in the presence of recombination, and higher overall recombination results in more and smaller blocks. But most recombinations do not lead to block boundaries. For instance, when the recombination rate approaches a level of \(10^{-5}/\text{kb}\) (which is thought to be a realistic average for humans; Kruglyak 1999), a block structure of LD still develops, even though many recombination events occur between all adjacent markers during the evolutionary history of the sample. The key observations are that high-LD areas are not characterized by reduced recombination frequency, and that block boundaries are not characterized by increased recombination frequency, not even by stochastic clustering of random recombination events (fig. 1).

It is straightforward to see that not all recombination events should affect block boundaries. Clearly, recombination will not affect LD values between two loci if it predates mutation events at these loci. Recombination also remains invisible if it involves two lineages that, looking backward in time, do not coalesce with any other lineage in the sample’s history before finding their common ancestor (Wiuf, Christensen, and Hein 2001). Furthermore, the impact on present-day LD of a recombination event is affected by subsequent recombinations that can revert allelic combinations to a previous state in part of the population. What is relevant is the degree to which a recombination event leaves a footprint of phylogenetic incompatibility in the present-day sample, meaning that observed haplotypes could not have arisen on one and the same phylogenetic tree in the absence of recurrent mutations. At a given genomic position, the total number of historic recombinations is not always a good predictor of such relevant recombinations (fig. 2).

The impact on LD block structure of timing of recombinations relative to mutation events is illustrated by the sensitivity of haplotype blocks to marker allele frequencies and historic population demographics (fig. 3). Biasing markers to exclude those with low minor allele frequencies results in a larger proportion of the genome contained in small blocks, as more recombination events have disrupted LD around high-frequency (old) mutations than around low-frequency (recent) mutations (Nordborg and Tavaré 2002). The magnitude of this effect can vary between populations. For instance, historic fluctuations in population size affect coalescence events differently than mutation events, as lineages that are ancestral to the sample coalesce faster when population size is small but do not mutate faster (Griffiths and Tavaré 1994). A severe bottleneck can therefore cause older branches in the genealogy to be short, allowing less opportunity to attract mutations. In such a case a larger proportion of randomly sampled mutations might be recent and would have experienced relatively little subsequent recombination, resulting in longer haplotype blocks. However, nonrandom sampling of old mutations only (assuming that those are still present abundantly in the sample) mitigates the effect and results in unaffected block structure. Note that in our models each genealogy carries the same number of mutations, irrespective of demographics, while the bottlenecks reduce the total length of the genealogy. This implies a higher overall mutation rate in the bottleneck models, but, alternatively, all models can be thought of as sampling different proportions of abundantly available polymorphisms.

It should also be noted that upscaling the coalescent model to the genealogy of long haplotypes with many markers challenges an assumption of the coalescent approximation that sample size is small relative to population size (Hudson 1990). Because of recombinations, the number of ancestors to the present-day sample can become large before coalescing in one common ancestor, and the assumption of small sample size may be violated, especially in bottleneck models. Although the results from our models are consistent with theoretical and empirical predictions that severe bottlenecks should increase the extent of LD (Kruglyak 1999), further theoretical work on the assumptions of the coalescent process is required to explore this issue.

Our simulations stress that the haplotype block structure can be weakly related to local historic recombination frequency, as only a small proportion of the recombinations might have a discernible effect on present-day LD. Observing haplotype blocks clearly does not justify the inference of recombination hot spots; additional evidence is needed to support their role in shaping the haplotype block structure.
Compelling evidence for recombination hot spots shaping LD patterns, albeit at a limited scale, is provided by sperm-typing studies that show concordance between population-level LD block boundaries and meiotic recombination clustering within single individuals (Jeffreys, Kauppi, and Neumann 2001; May et al. 2002; Kauppi, Sajantila, and Jeffreys 2003). Simulation studies also suggest a role for recombination hot spots, as some characteristics of human LD block patterns, such as incidental very long blocks or the degree of genomic block coverage, are not easily reproduced under uniform recombination (Wang et al. 2002; Wall and Pritchard 2003a; Zhang et al. 2003; but see Phillips et al. 2003). On the other hand, our results suggest that local LD patterns may be more affected by the relative timing of historic recombination and mutation events in their genomic region than by the total number of recombinations, and infrequent recombination events can easily result in block boundaries in areas of reduced recombination (Wall and Pritchard 2003b). Furthermore, the human LD block structure appears sensitive to population demographic history in a way that is consistent with stochastic determination of block boundaries under uniform recombination. For instance, African samples consistently show less extensive haplotype blocks than samples from the more recently founded European population (Gabriel et al. 2002; Wall and Pritchard 2003b). In some studies block boundaries have been reported to correspond well across populations (Gabriel et al. 2002; Kauppi, Sajantila, and Jeffreys 2003), consistent with the idea that recombination hot spots shape the haplotype block structure. However, these studies considered mainly old mutations (that is, with high minimum allele frequencies) that predate some or all of the population separations. This strategy can detect shared block structures that derive from shared population history, but is not efficient in detecting population-specific block structures that developed after the populations separated (Wang et al. 2002; Zhang et al. 2003).

Our simulations assume selective neutrality of mutations, and therefore they do not capture the effect that selection has on LD patterns. Selection can affect haplotype block structure: a beneficial mutation can sweep through the population together with linked alleles of the haplotype on which it arose (the hitchhiking effect; Maynard Smith and Haigh 1974), creating an area of reduced haplotype diversity and increased LD (that is, a haplotype block) around the selected locus (Sabeti et al. 2002; Wootton et al. 2002; Palaisa et al. 2004). The affected area can be large initially (Sabeti et al. 2002), but it may erode rapidly as a result of recombination (Przeworski 2002). Recent work in coalescent theory is aimed at exploring the spatial effects of such selective sweeps along recombining chromosomes (e.g., Kim and Stephan 2002; Kim and Nielsen 2004), which can improve our understanding of the effects of selective sweeps on haplotype block patterns. It is relevant to note that LD blocks that arise due to selective sweeps will develop in the absence of recombination hot spots. As with the...
stochastically developing haplotype blocks in our simulations, they are affected by recombination, but their structure does not necessarily reflect intrinsic spatial variation in recombination intensity. Rather, they are thought to vary between populations according to population-specific selection pressures (Storz, Payseur, and Nachman 2004).

Conclusions

Given our current knowledge, it is evident that punctuated recombination plays a role in shaping human haplotype block patterns, but it remains to be established how large that role is relative to the effects of stochastic interplay between random recombinations, mutation, and coalescent events. Our simulations stress the potential of stochastic events to shape haplotype blocks. Empirical evidence for the impact of rare and stochastic historic events on haplotype block structure will be difficult to obtain directly. However, careful examination of present-day recombination patterns (for instance in sperm-typing studies) can show whether spatial variation in recombination intensity, as observed in single individuals, corresponds to population-wide haplotype block structure. As simulation studies like ours indicate that the role of stochastic events might be large, it seems that considerable evidence from such empirical studies would be required to demonstrate that recombination hot spots are the main determinants of the haplotype block structure.

Acknowledgments

We thank Lauren McIntyre for many helpful suggestions and Dan Noland for programming assistance. We are deeply grateful to the late David P. Moffett, whose timely interest and provision of programming support helped make this project possible. This work was supported by the National Science Foundation, DBI-9904704.

Literature Cited


Sharon Wolf, Associate Editor

Accepted November 22, 2004