The Structure of Haplotype Blocks in the Human Genome

Article *in* Science · July 2002

DOI: 10.1126/science.1069424 · Source: PubMed

18 authors, including:

Stephen Schaffner
Broad Institute of MIT and Harvard
345 PUBLICATIONS **59,250** CITATIONS
SEE PROFILE

Matthew DeFelice
Broad Institute of MIT and Harvard
22 PUBLICATIONS **18,476** CITATIONS
SEE PROFILE

Adebowale Adeyemo
National Institutes of Health
592 PUBLICATIONS **27,800** CITATIONS
SEE PROFILE

Sciencexpress Report

The Structure of Haplotype Blocks in the Human Genome

Stacey B. Gabriel,[1] Stephen F. Schaffner,[1] Huy Nguyen,[1] Jamie M. Moore,[1] Jessica Roy,[1] Brendan Blumenstiel,[1] John Higgins,[1] Matthew DeFelice,[1] Amy Lochner,[1] Maura Faggart,[1] Shau Neen Liu-Cordero,[1,2] Charles Rotimi,[3] Adebowale Adeyemo,[4] Richard Cooper,[5] Ryk Ward,[6] Eric S. Lander,[1,2] Mark J. Daly,[1] David Altshuler[1,7]*

[1]Whitehead/ MIT Center for Genome Research, Cambridge, MA 02139, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA. [3]National Human Genome Center, Howard University, Washington, DC 20059, USA. [4]Department of Pediatrics, College of Medicine, University of Ibadan, Ibadan, Nigeria. [5]Department of Preventive Medicine and Epidemiology, Loyola University Medical School, Maywood, IL 60143, USA. [6]Institute of Biological Anthropology, University of Oxford, Oxford, England OX2 6QS. [7]Departments of Genetics and Medicine, Harvard Medical School; Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA.

*To whom correspondence should be addressed. E-mail: altshuler@molbio.mgh.harvard.edu

Haplotype-based methods offer a powerful approach to disease gene mapping, based on association between causal mutations and the ancestral haplotypes on which they arose. We characterized haplotype patterns across 51 autosomal regions (spanning 13Mb of the human genome) in samples from Africa, Europe and Asia. We show that the human genome can be parsed objectively into haplotype blocks: sizeable regions over which there is little evidence for historical recombination, and within which only a few common haplotypes are observed. The boundaries of blocks and specific haplotypes they contain are highly correlated across populations. We demonstrate that such haplotype frameworks provide substantial statistical power in association studies of common genetic variation across each region. Our results provide a foundation for the construction of a haplotype map of the human genome, facilitating comprehensive genetic association studies of human disease.

Variation in the human genome sequence plays a powerful but poorly understood role in the etiology of common medical conditions. As the vast majority of heterozygosity in the human population is attributable to common variants, and since the evolutionary history of common human diseases (which determined the allele spectrum for causal alleles) is not yet known, one promising approach is to comprehensively test common genetic variation for association to medical conditions(1-3). This is increasingly practical, with four million (4, 5) of the estimated ten million (6) common single nucleotide polymorphisms (SNPs) already known.

In designing and interpreting association studies of genotype and phenotype, it is necessary to understand the structure of haplotypes in the human genome. Haplotypes are the particular combinations of variant alleles observed in a population. When a new mutation arises, it does so on a specific chromosomal haplotype. The association between each mutant allele and its ancestral haplotype is disrupted only by mutation and recombination in subsequent generations. Thus, it should be possible to track each variant allele in the population by identifying (through the use of anonymous genetic markers) the particular ancestral segment on which it arose. Haplotype methods have contributed to the identification of genes for Mendelian diseases (7-9), and recently, common, complex disorders (10-12). The general properties of haplotypes in the human genome, however, have remained unclear.

Many studies have examined allelic associations (also termed "linkage disequilibrium") across one or a few gene regions. These studies have generally concluded that linkage disequilibrium is extremely variable both within and among loci and populations (reviewed in (13-15). Recently, examination of a higher density of markers over contiguous regions (16-18) suggested a surprisingly simple pattern: blocks of variable length over which only a few common haplotypes are observed, punctuated by sites at which recombination could be inferred in the history of the sample. In one segment of the HLA, it has been directly demonstrated that "hotspots" of meiotic recombination coincided with boundaries between such blocks (17). These studies suggested a model for human haplotype structure, but left many questions unanswered. First, how much of the human genome exists in such blocks, and what is the size and diversity of haplotypes within blocks? Second, to what extent do these characteristics vary across population samples? Third, can haplotype patterns be parsed using only common SNPs sampled from the population, or will the pattern only emerge after complete resequencing (19)? Fourth, how completely does such a haplotype framework capture common sequence variation within each region?

To determine the general structure of human haplotypes, we selected 54 autosomal regions, each with an average size of 250,000 bp, spanning in total 13.4 Mb (≈0.4%) of human genome. Regions were selected according to two criteria: that they be evenly spaced throughout the genome, and that they contain an average density (in a core region of 150kb) of one candidate SNP discovered by The SNP Consortium every 2kb (20) (Supplemental table1). Genotyping was performed by primer extension of multiplex products with detection by MALDI-TOF mass spectroscopy(21) (22). Each SNP was genotyped in 275 individuals (400 independent chromosomes) sampled from four population groups: 30 parent-offspring trios (90 individuals) from Nigeria (Yoruba), 93 members of 12 multigenerational pedigrees of European ancestry (Utah CEPH), 42 unrelated individuals of Japanese and Chinese origin, and 50 unrelated African Americans.

We designed assays to 4,532 candidate SNPs of which 3,738 (82%) were successfully genotyped (23) (24). Three of the 54 regions were withheld from further analysis due to

inconsistencies in genome assembly and/or evidence for a closely related paralogous region (making locus-specific PCR difficult). In the remaining 51 regions, accuracy of genotype calls was empirically assessed as ≈99.6 % (25). We note that a very low rate of genotyping error is absolutely necessary for studies of multi-marker haplotypes: even a modest error rate creates the appearance of "rare variant' haplotypes that do not exist in nature.

Of candidate TSC SNPs successfully assayed, 89% were verified to be polymorphic in one or more populations. The proportion polymorphic in each sample varied from 70% (Asian) to 86% (African American) (Figure 1a). Although the majority of SNPs (59%) were observed in all four populations, there are dramatic differences in the allele frequencies of individual SNPs across samples (Fig. S1a-e) consistent with prior estimates of population differentiation and origin (26).

If haplotype blocks represent regions inherited without significant recombination in the ancestors of the current population, then a biological basis for defining haplotype blocks is to examine patterns of recombination across each region. The history of recombination between a pair of SNPs can be estimated using the normalized measure of allelic association D' (16, 27). Since D' values are known to fluctuate upwards when small number of samples or rare alleles are examined, we relied on confidence bounds on D' rather than point estimates (28). We define pairs to be in "strong LD" if the one-sided upper 95% confidence bound is > 0.98 (that is, consistent with no historical recombination) and the lower bound is above 0.7 (29). Conversely, we term "strong evidence for historical recombination" pairs for which the upper confidence bound on D' is less than 0.9. On average, 87% of all pairs of markers with minor allele frequency > 0.2 fell into one of these two categories (termed "informative" marker pairs). This method should be robust to study-specific differences in the frequencies of SNPs and sample sizes examined, since it relies on those pairs for which narrow confidence intervals (that is, precise estimates) have been obtained.

When this definition is applied to pairs of markers separated by less than 1,000 bp, a small fraction of informative pairs show strong evidence of historical recombination (Fig. 1B): 14-18% in the Yoruban and African American samples, and 3-6% in the European and Asian samples. In the Yoruban and African American samples, the proportion of pairs displaying evidence for historical recombination rises rapidly with distance, increasing to 50% at a separation of ≈8kb. In the European and Asian samples, in contrast, the fraction of pairs showing strong evidence for recombination rises to 50% at 22kb. These differences in LD among populations are likely attributable to differences in demographic history(30), since the biological determinants of LD (rates of recombination, mutation, gene conversion) are expected to be constant across groups. The data show that LD extends to a similar and long extent in Asian as well as European samples, and that African American samples show very similar patterns to those observed in the Yoruban population.

The spatial distribution of D' values across each region (for example, see fig. S2) demonstrated clusters of markers over which strong evidence of historical recombination was minimal. We defined a haplotype block as a region over which a very small proportion (<5%) of comparisons among informative SNP pairs show strong evidence of historical recombination. (We allow for 5% because many forces other than recombination (both biological and artifactual) can disrupt haplotype patterns: recurrent mutation, gene conversion, errors of genome assembly or genotyping.) We implemented this definition in two ways. Where many markers were sampled, we simply counted the proportion of pairs with strong evidence of historical recombination. Over much of our survey, however, we observed regions in which all of the informative markers showed strong evidence of linkage disequilibrium, but the number of comparisons was insufficient to confidently conclude (simply by counting) that the proportion of such pairs was >95%. By systematically sampling the entire dataset, however, we found that information from as few as two or three markers could suffice to identify regions as blocks (Fig. 2A, 2B). These criteria (31) allowed us to define blocks even where the marker coverage is less complete.

Armed with these criteria, we systematically examined the dataset for haplotype blocks, identifying a total of 928 blocks in the four populations samples. Within blocks, independent measures of pairwise linkage disequilibrium did not decline substantially with distance (Fig. 2C). The minimum span of the blocks (measured as the interval between the flanking markers used to define them) averaged 9 kb in the Yoruban and African American samples, and 18kb in the European and Asian samples. The size of each block varied dramatically, however: from <1kb to 94 kb in the African American and Yoruban samples and from <1kb to 173kb in the European and Asian samples. While most of the blocks were small (Fig. 3A), most of the sequence spanned by blocks was in large blocks (Fig. 3B).

Our survey consisted of randomly spaced markers (based on the public map), averaging one marker (with frequency > 0.1) every 7.8 kb across the regions surveyed. The partial information leads to two biases in block detection. First, in regions in which we had few markers, we are less likely to detect small blocks. Conversely, identified blocks will typically extend some distance beyond the randomly spaced markers that happen to fall within their boundaries. To estimate the true distribution of block sizes, we performed computer simulations in which block sizes were exponentially distributed (with a specified average size), and markers were randomly spaced (with a mean spacing equal of one every 7.8 kb). These simulations provided a good fit to the observed data when the mean size of blocks was estimated to be 11 kb in the Yoruban and African American samples, and 22 kb in the European and Asian samples (Table 1; (32)). This corresponds to an N50 size 22kb in the Yoruban and African American samples, and of 44kb in the European and Asian populations. (The N50 size is defined as the length *x* such that 50% of the genome lies in blocks of *x* or longer.) In addition, the model predicts that the proportion of the human genome spanned by blocks of 10kb or larger is 65% in the Yoruban and African American samples, and 85% in the European and Asian samples.

We next examined haplotype diversity within blocks. We note that our block definition, unlike one previously proposed(18), is based on recombination, and thus does not require low haplotype diversity. Nevertheless, within regions with scant evidence for historical recombination, we observe only three to five common (>5%) haplotypes in each of population samples (Fig. 3C). As few as 6-8 randomly chosen common markers suffice to identify these common haplotypes: the number of haplotype reached a plateau with 6-8 common markers, with little evidence for the discovery of additional common haplotypes if up to 17 markers are

included (Fig. 3C). Thus, low haplotype diversity is not simply an artifact of having examined only a small number of markers, but is a true feature of regions with low rates of historical recombination. Haplotype diversity was greatest in the Yoruban and African-American samples, with an average of 5.0 common haplotypes observed. Lower diversity was observed in the European samples (4.2 common haplotypes), and the smallest number of common haplotypes (3.5) was observed in the Asian samples. Critically, even where many markers are examined, these few common haplotypes explained the vast majority (≈90%) of all chromosomes in each population sample (Fig. 3D). (33)(34)

To compare block boundaries across different populations, we examined adjacent pairs of SNPs successfully assayed in at least two populations. In each population, we asked whether the pair were assigned to a single block, or showed strong evidence of historical recombination. A SNP pair was termed concordant if the assignment was the same in both populations, and discordant if the assignments disagreed (35). We found the great majority of SNP pairs (77% - 95%, depending on the population comparison) were concordant across population samples (Fig. 4A-D). Moreover, where discordance across populations was observed, it was nearly always due to pairs displaying strong evidence of historical recombination in the Yoruban and African American samples, but not in the European and Asian samples (Fig. 4A-D).

We compared the specific haplotypes observed across the European, Asian, and Yoruban (African) samples. To ensure that haplotype diversity was well defined in this comparison, we considered only those blocks in which six or more polymorphic markers were obtained (36). Each single population sample contained 3.1 – 4.9 haplotypes with a frequency >5%. The union of these sets, however, contained only 5.3 haplotypes (Fig. 4E). That is, the specific haplotypes observed in each group were remarkably similar: 51% (2.7) were identified in all three populations, and 72% in two of the three groups. Moreover, of the 28% of haplotypes found in a only one population sample, nearly all (90%) were found in the Yoruban sample. The similarity in haplotype identities across the European and Asian samples is striking, with an average of only 0.1 haplotypes per block that were unique to either population sample.

The comparison across populations of SNP polymorphism (Fig. 1A, Fig. S1A-D) recombinant sites (Fig. 4A-D) and haplotypes (Fig. 4E) are supportive of a single "out of Africa" origin (37, 38) for both the European and Asian samples. The data suggest a significant bottleneck in the ancestry of these samples, with only a subset of the diversity (of SNPs, of haplotypes, and of recombinant chromosomes) in Africa found in the two non-African populations(30, 39-42). Since bottlenecks preferentially effect lower-frequency alleles, this model predicts that the alleles (haplotypes and recombinant chromosomes) present only in the African samples would have lower allele frequencies in Africa than are pan-ethnic alleles, and our data support this hypothesis (43).

The major attraction of haplotype methods is the idea that common haplotypes capture most of the genetic variation across sizeable regions, and that these haplotypes (and the undiscovered variants they contain) can be tested using a small number of haplotype tag SNPs ("htSNPs") (16, 18, 19, 44). A number of reports (44-46), however, have suggested that many SNPs fail to conform to the underlying haplotype structure, and would be missed by haplotype based approaches.

To examine this question empirically, we defined a framework of haplotype blocks using a randomly-selected subset of our data (requiring a minimum of 6 markers per block), and examined the correlation coefficient (r^2) between these haplotypes and an additional set of SNPs (not used to define the blocks) within their span. These additional SNPs were meant to model the undiscovered variation in each region that one would hope to track using a haplotype approach. We found that the average maximal r^2 value between each additional SNP and the haplotype framework was high: 0.67 to 0.87 in the four population samples. That is, for the average untested marker, only a small increase in sample size (15-50%) would be needed using a haplotype-based study rather than discovering and testing that SNP directly. Moreover, we find that within blocks, a large majority (77-93%) of all untested markers showed r^2 values greater than 0.5 to the framework haplotypes (47). These results demonstrate that haplotype blocks can be used to study association to the vast majority of variants within each region with little loss of statistical power.

Our results show that haplotype blocks can be reliably identified by genotyping a sample of common markers within their span; that is, without complete resequencing. To have confidence that a region is a block, however, requires typing a high density of polymorphic markers in a sufficiently large sample to confidently parse the patterns of historical recombination across the region. Our data provide strong evidence that most of the human genome is contained in blocks of substantial size: we estimate that half of the human genome exists in blocks of 22kb or larger in African and African American samples, and in blocks of 44kb or larger in European and Asian samples. Within each block, a very small number of common haplotypes (three to five) typically capture ≈90% of all chromosomes in each population. Both the boundaries of blocks and the specific haplotypes observed are shared to a remarkable extent across populations, with the main variation being a subset of alleles (haplotypes and recombinant forms) that are observed only in samples with more recent African ancestry. Finally, blocks defined with a small number of common markers do a quite comprehensive job of capturing the common variation across each locus.

Our results provide a methodological and quantitative foundation for the construction of a haplotype map of the human genome using common SNP markers. Although the patterns are simpler and haplotypes longer than some had predicted, we note that our results suggest that very dense SNP coverage will be needed to complete such a map. With an average block size of 11 to 22kb and three to five haplotypes per block, our data suggest that fully powered haplotype association studies could ultimately require as many as 300,000-1,000,000 well-chosen htSNPs (in non-African and African samples, respectively). This number represents an upper limit, however: there is often significant linkage disequilibrium between adjacent blocks (data not shown), allowing fewer markers to be used without loss of power. It will likely be productive to perform initial haplotype mapping in populations whose history contains one or more bottlenecks, as longer range LD may make initial localization more efficient and favorable. Conversely, populations with shorter-range LD and greater haplotype diversity may be offer advantages for fine mapping. In suggesting that block boundaries and common haplotypes are largely shared across populations, our data suggest that many common disease alleles can be studied — and will likely be broadly relevant — across human populations. In the future,

comprehensive analysis of human haplotype structure promises new insights into the origin of human populations, the forces that shape genetic diversity, and the population basis of disease.

References and Notes

1. E. S. Lander, *Science* **274**, 536-9 (1996).
2. F. S. Collins, M. S. Guyer, A. Charkravarti, *Science* **278**, 1580-1 (1997).
3. N. Risch, K. Merikangas, *Science* **273**, 1516-7 (1996).
4. R. Sachidanandam *et al.*, *Nature* **409**, 928-33. (2001).
5. J. C. Venter *et al.*, *Science* **291**, 1304-51. (2001).
6. L. Kruglyak, D. A. Nickerson, *Nat Genet* **27**, 234-6. (2001).
7. E. G. Puffenberger *et al.*, *Cell* **79**, 1257-66. (1994).
8. B. Kerem *et al.*, *Science* **245**, 1073-80. (1989).
9. J. Hastbacka *et al.*, *Nat Genet* **2**, 204-11 (1992).
10. J. D. Rioux *et al.*, *Nat Genet* **29**, 223-8. (2001).
11. J. P. Hugot *et al.*, *Nature* **411**, 599-603. (2001).
12. Y. Ogura *et al.*, *Nature* **411**, 603-6. (2001).
13. J. K. Pritchard, M. Przeworski, *Am J Hum Genet* **69**, 1-14. (2001).
14. L. B. Jorde, *Genome Res* **10**, 1435-44. (2000).
15. M. Boehnke, *Nat Genet* **25**, 246-7 (2000).
16. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, E. S. Lander, *Nat Genet* **29**, 229-232 (2001).
17. A. J. Jeffreys, L. Kauppi, R. Neumann, *Nat Genet* **29**, 217-22. (2001).
18. N. Patil *et al.*, *Science* **294**, 1719-23. (2001).
19. G. C. Johnson *et al.*, *Nat Genet* **29**, 233-7. (2001).
20. Materials and methods are available as supporting material on Science Online
21. Materials and methods are available as supporting material on Science Online
22. K. Tang *et al.*, *Proc Natl Acad Sci U S A* **96**, 10016-20. (1999).
23. Materials and methods are available as supporting material on Science Online
24. While 82% of assays were successful in at least one population, genotyping success rates in each population range from 72% to 79%. The difference between these numbers is due to a low rate of laboratory failure in each attempt.
25. Materials and methods are available as supporting material on Science Online
26. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The history and geography of human genes* (Princeton University Press, Princeton, NJ, 1994).
27. R. C. Lewontin, *Genetics* **49**, 49-67 (1964).
28. Materials and methods are available as supporting material on Science Online
29. An upper confidence bound of 0.98 was used instead of 1.0 because even a single observation of a fourth haplotype makes it is mathematically impossible for D' to be consistent with a value of 1.0, even though the confidence interval could be arbitrarily close to 1.0
30. D. E. Reich *et al.*, *Nature* **411**, 199-204. (2001).
31. Materials and methods are available as supporting material on Science Online
32. As a further test of the model, we simulated the proportion of pairs at a fixed distance (5kb) that should show evidence of crossing block boundaries (that is, show strong evidence of historical recombination). The model predicts these proportions to be 47% (Yoruban and African American samples), and 27% (European and Asian samples). In the empirical data, we observe 42% and 23%, similar to these predictions.
33. A low rate of genotyping error is critical to obtaining an accurate measure of haplotype diversity and the proportion in common haplotypes. Even a modest (1-2%) genotyping error will create a substantial number of false rare haplotypes: for example, with a 10 marker haplotype and a 2% error rate, 18% of chromosomes will contain at least one error, and thus not match the few common haplotypes.
34. Within blocks, the common haplotypes showed little evidence for historical recombination. For example, we performed the four gamete test using SNPs drawn only from haplotypes with frequency 5% or higher in each block. In only 5% of blocks was a one or more violation to the four gamete test observed.
35. To maximize power, these comparisons were made only for SNP pairs spaced five to ten kilobases apart. At shorter distances, nearly all SNP pairs are in a single block, and at greater distances, most SNP pairs are in different blocks.
36. Blocks and haplotypes were identified separately in each population sample, and the results compared for those blocks that were physically overlapping in all three samples.
37. R. L. Cann, W. M. Brown, A. C. Wilson, *Genetics* **106**, 479-99. (1984).
38. C. B. Stringer, P. Andrews, *Science* **239**, 1263-8. (1988).
39. D. E. Reich, D. B. Goldstein, *Proc Natl Acad Sci U S A* **95**, 8119-23 (1998).
40. M. Ingman, H. Kaessmann, S. Paabo, U. Gyllensten, *Nature* **408**, 708-13. (2000).
41. S. A. Tishkoff *et al.*, *Science* **271**, 1380-7. (1996).
42. S. A. Tishkoff *et al.*, *Am J Hum Genet* **67**, 901-25. (2000).
43. We examined SNP pairs that were in different blocks in the Yoruban samples but in a single block in the European sample. Such pairs had higher D' values in the Yoruban sample (D' = 0.46) than did pairs found in different blocks in both population samples (D' = 0.28). The average frequency of all haplotypes in the Yoruban population was 0.21, while those that were found only in the Yoruban sample (but not in the European and Asian samples) had a mean frequency of 0.16.
44. A. G. Clark *et al.*, *Am J Hum Genet* **63**, 595-612 (1998).
45. A. R. Templeton *et al.*, *Am J Hum Genet* **66**, 69-83 (2000).
46. S. M. Fullerton *et al.*, *Am J Hum Genet* **67**, 881-900 (2000).
47. The small fraction of SNPs that show r^2 values < 0.5 could be attributable to a range of causes: branches of the gene tree not defined with the number of markers employed, gene conversion events or recurrent mutations. We note that errors in genotyping or map position decrease (but cannot increase) the value of r^2.
48. Materials and methods are available as supporting material on Science Online

49. This work was supported by a grant to DA from The SNP Consortium. We thank members of the Program in Medical and Population Genetics at the Whitehead/MIT Center for Genome Research for helpful discussion; particularly Joel Hirschhorn, David Reich and Nick Patterson. DA is a Charles E. Culpeper Scholar of the Rockefeller Brothers Fund, and a Burroughs Welcome Fund Clinical Scholar in Translational Research.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1069424/DC1
Materials and Methods

figs. S1, S2, S3
table S1

28 December 2001; accepted 13 May 2002

Published online 23 May 2002;
<zdoi;10.1126/science.1069424>

Include this information when citing this paper.

Fig.1. (A) Normalized allele frequency of candidate SNPs. The distribution is normalized to a constant number of chromosomes (n=64 randomly sampled) from the European, African-American, Asian, and Yoroban samples. Of candidate SNPs assayed in all four populations, both predicted alleles were observed in 89% of cases. (B) Assessment of pairwise linkage disequilibrium across populations. The proportion of informative SNP pairs that display strong evidence for recombination (see text) is plotted at various intermarker distances. Between 9,860 and 13,980 SNP pairs were examined in each sample.

Fig. 2. (A,B) Scaffold analysis of Yoruban and African American (A), and European and Asian (B) samples. The y-axis indicates the fraction of independent, informative marker pairs (within each region) displaying strong evidence for recombination. The x-axis indicates the distance between the outermost marker pair defining the region. The three lines represent the distribution of LD for of all pairs (without any filtering for the LD of flanking markers), and for regions meeting the empirically derived two and three marker criteria (48). (C) Relationship of linkage disequilibrium to physical distance within haplotype blocks, as assessed by the mean value of the correlation coefficient (r2) and the mean value of D'. The marker pairs in this figure were not used to define the region as a block, and thus represent an unbiased estimation of the relationship between LD and distance within a block.

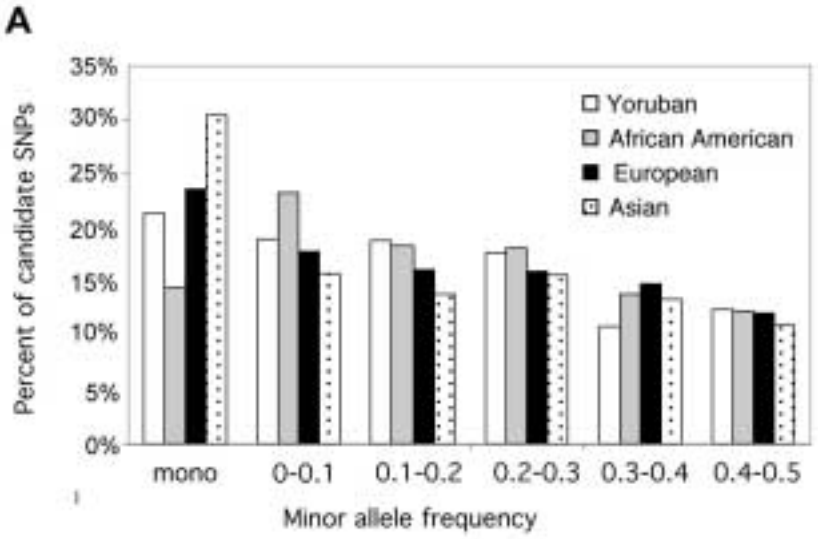
Fig. 3. Block characteristics across populations. (A) Size (kb) distribution of all haplotype blocks found in the analysis. (B) Proportion of all genome sequence spanned by blocks, binned according to the size of each block. (C,D) Summary of haplotype diversity across all blocks. The number of common (≥5%) haplotypes per blocks (C) and fraction of all chromosomes representing a perfect match to one of these common haplotypes (D) is plotted as a function of the number of markers typed in each block.

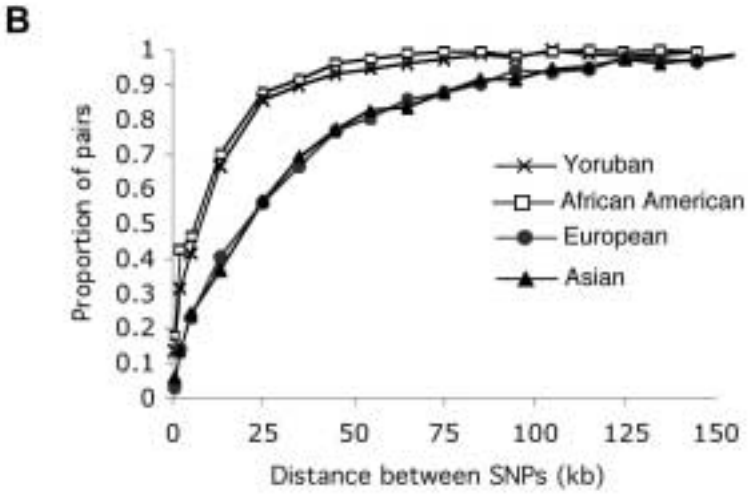
Fig. 4. Comparison of blocks across population samples. (A-D) Concordance of block assignments for adjacent SNP pairs, compared across populations. In each plot, the white bars represent the fraction of concordant SNP pairs, and the black bars the proportion of discordant SNP pairs. Population samples are abbreviated as EU, European sample; AS, Asian sample; AA, African American sample; YR, Yoruban sample. (E) Distribution of haplotypes across populations.

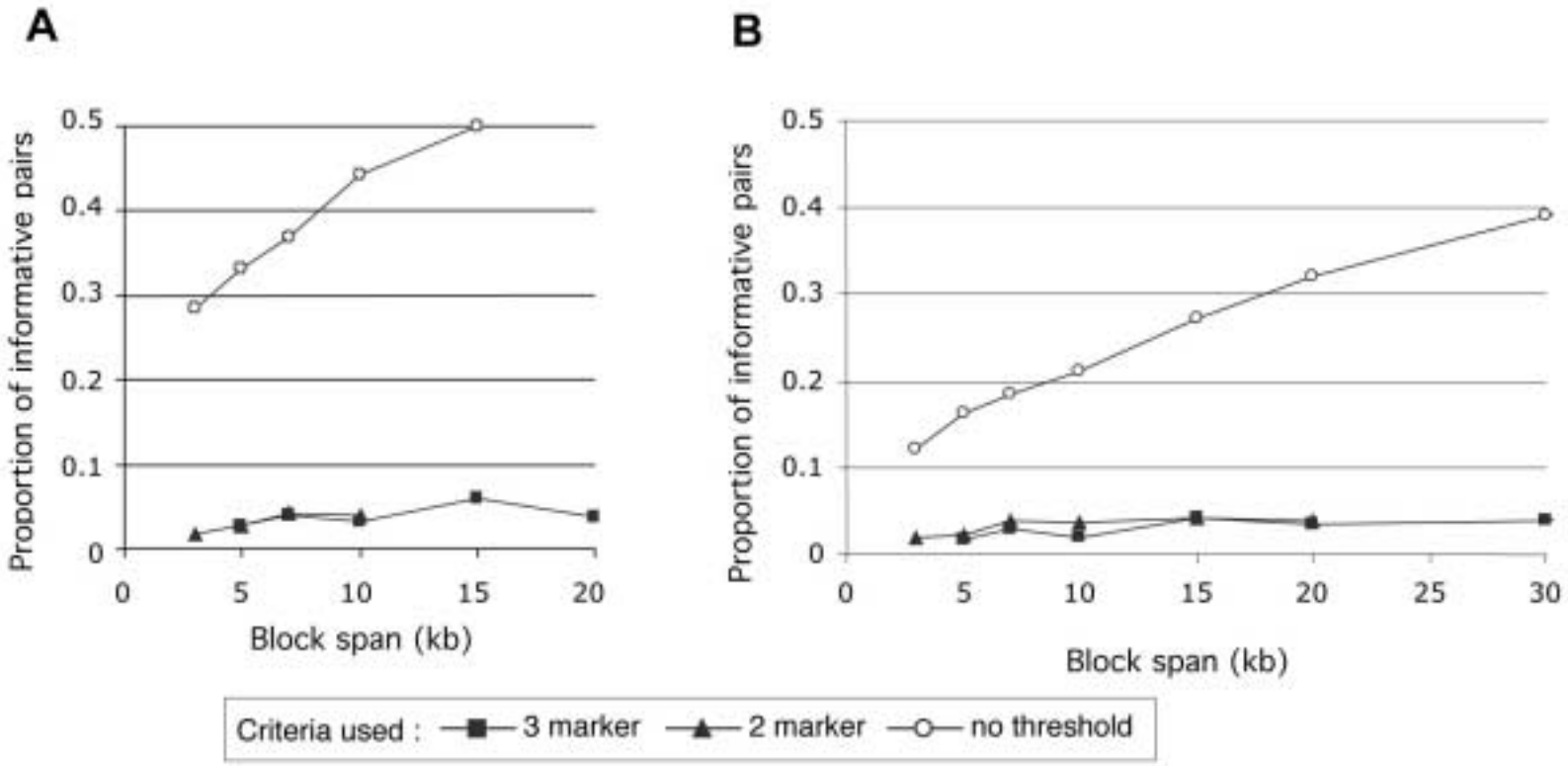
Table 1. Observed and predicted proportion of sequence found in haplotype blocks. Model is based on the best fit to the observed data, and assumes randomly spaced markers with an average density of one every 7.8 kb, and block span an exponentially distributed random variable with a mean size in European sample of 22 kb and of 11 kb in the Yoruban sample. In the model, block boundaries of 2 kb in length are assumed [Jeffreys, 2001 #1042]. Although the observed and predicted values were not statistically significantly different (data not shown), we note that both models show a trend towards underestimating the incidence of short blocks (0-5kb).

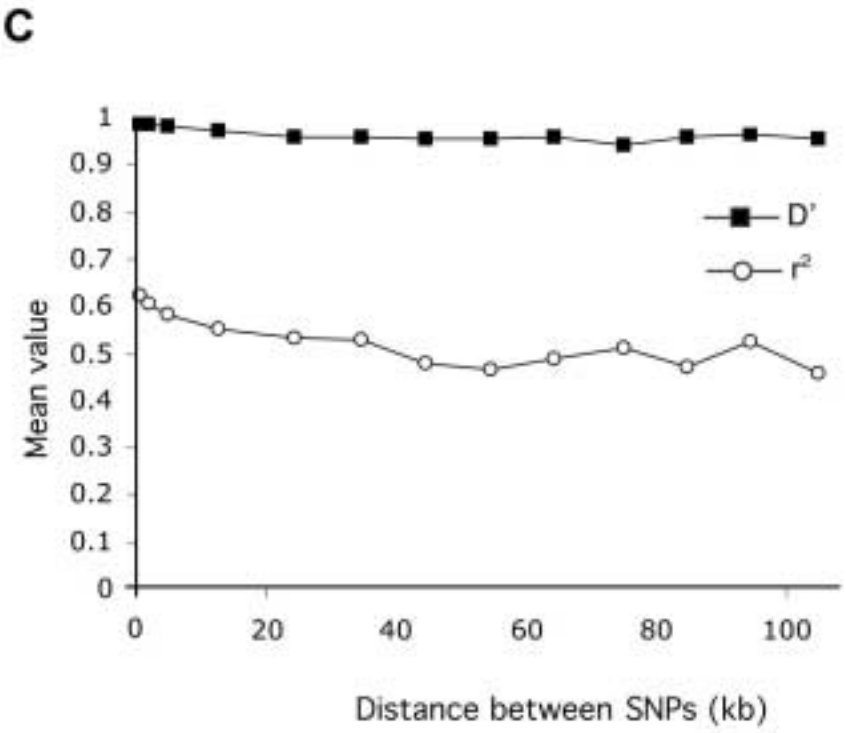
Table 1 : Model versus oberseved distribution of block sizes

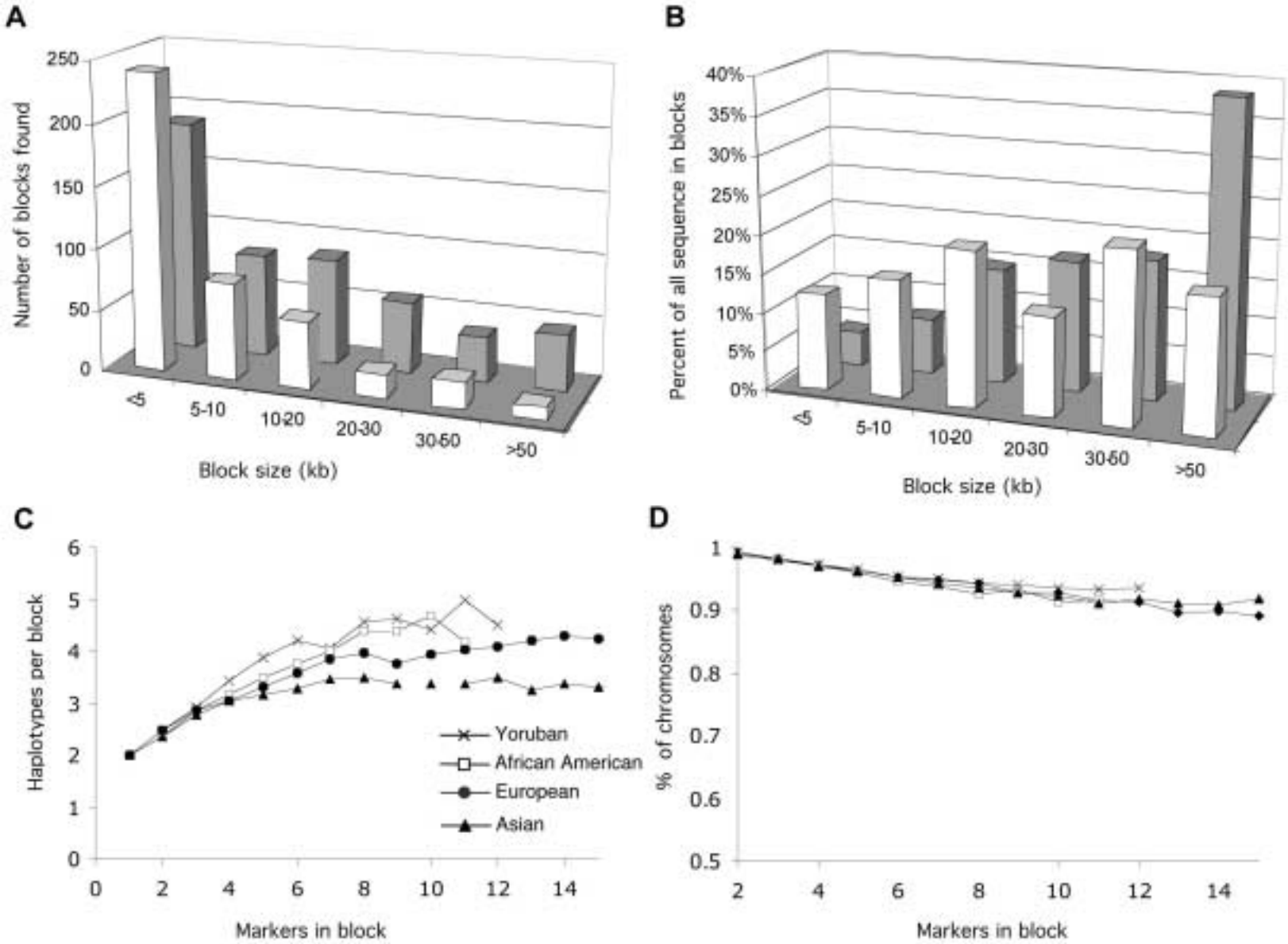
	African sample		Non-African sample	
block size	observed % of spanned sequence	predicted % of spanned sequence	observed % of spanned sequence	predicted % of spanned sequence
0-5	12.4	6.3	4.4	1.8
5-10	15.3	15.1	7.4	5.2
10-20	20	31.5	14.9	15.2
20-30	12.8	21.8	16.6	16.6
30-50	22.2	19.1	18	26.9
>50	17.4	6.3	38.7	34.2


A
35%
30%
25%
20%
15%
10%
5%
0%
Percent of candidate SNPs
Yoruban
African American
European
Asian
mono
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
Minor allele frequency


B
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion of pairs
Yoruban
African American
European
Asian
0
25
50
75
100
125
150
Distance between SNPs (kb)


A
B
Proportion of informative pairs
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
Block span (kb)
Criteria used :
3 marker
2 marker
no threshold


C
Mean value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
D'
r²
0
20
40
60
80
100
Distance between SNPs (kb)


A
Number of blocks found
250
200
150
100
50
0
<5
5-10
10-20
20-30
30-50
>50
Block size (kb)
B
Percent of all sequence in blocks
40%
35%
30%
25%
20%
15%
10%
5%
0%
<5
5-10
10-20
20-30
30-50
>50
Block size (kb)
C
Haplotypes per block
6
5
4
3
2
1
0
0
2
4
6
8
10
12
14
Markers in block
Yoruban
African American
European
Asian
D
% of chromosomes
1
0.9
0.8
0.7
0.6
0.5
2
4
6
8
10
12
14
Markers in block