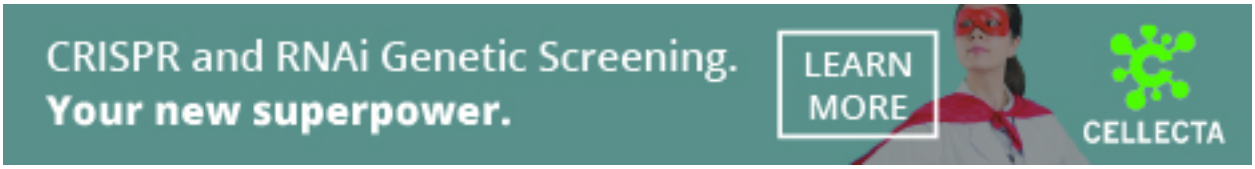
Human Haplotype Block Sizes Are Negatively Correlated With Recombination Rates

Tiffany A. Greenwood, Brinda K. Rana and Nicholas J. Schork

Genome Res. 2004 14: 1358-1361
Access the most recent version at doi:10.1101/gr.1540404



Letter

Human Haplotype Block Sizes Are Negatively Correlated With Recombination Rates

Tiffany A. Greenwood, Brinda K. Rana, and Nicholas J. Schork[1]

Polymorphism Research Laboratory, Department of Psychiatry, University of California at San Diego, La Jolla, California 92093, USA

The International Haplotype Map ("HapMap") Project is motivated, in part, by the belief that the organization of the human genome, the mechanics of recombination, and the population-level behavior of alleles at adjacent loci should allow researchers to parse the genome into small segments, or "blocks," that show strong linkage disequilibrium (LD) between alleles at loci within those segments. The discovery and evidence for these blocks is to be based solely on the observed LD strength and patterns between alleles at adjacent loci throughout the genome. Although there are many factors that contribute to LD strength, we assessed the correlation between block structure, in terms of length and percentage of the genome assembled into blocks within a region, and recombination rate obtained from two independent sources. We found evidence of a striking negative correlation between the average recombination rate and average block length, suggesting that recombination rate is a strong contributor to haplotype block structure within the genome. We discuss the potential implications of this negative correlation in the context of the organization, properties, and potential ubiquity of a block-like structure in the human genome.

[Supplemental material is available online at www.genome.org.]

Many research groups have begun to consider and assess evidence that the genome can be partitioned into haplotype blocks based on the patterns of linkage disequilibrium (LD) exhibited by alleles at adjacent loci (http://www.hapmap.org; Daly et al. 2001; Johnson et al. 2001; Olivier et al. 2001; Patil et al. 2001; Dawson et al. 2002; Gabriel et al. 2002; Stumpf 2002). Because LD patterns are used to identify these blocks, and because there is no consensus regarding the best or most optimal statistical method to uncover blocks via LD patterns, it is unclear whether the blocks uncovered by one group will be consistent with the blocks uncovered by another group, even within the same racial or ethnic group. In addition, as one very important ancillary goal of the HapMap Project is to identify the minimal number of polymorphic sites that one needs to consider for genotyping in order to reconstruct the possible haplotypes within a block, it is important to consider the "ubiquity" of any proposed block structure.

The ubiquity of the haplotype block structure of the genome is an important question to address for another reason: There are a great number of factors that contribute to LD strength (i.e., the phenomenon used to define blocks) within and across populations, so it is important to consider which factors contribute to block structure and which do not. For example, founding population size, the number of generations that have elapsed since the founding of a population, population expansion rate, immigration rates, and variation in the number of offspring per mating all influence LD strength, but are factors that are unique to different populations (see, e.g., Lewontin 1974). However, factors involved in the organization and meiotic recombination of chromosomes, such as recombination hot-spots, genomic sites for heavy mutation and gene conversion, and the orientation of 'neutral' and selectively important (i.e., gene-rich) regions of the genome also influence LD strength, and may operate similarly and ubiquitously across all human populations. With this in mind, Zhang et al. (2003) argued that genetic drift and other stochastic processes known to influence LD in populations can create a block-like structure in LD patterns across the genome that is somewhat independent of recombination rates. The implication of the Zhang et al. (2003) study is that if these population-level factors dictate block-like LD patterns in the genome, then a block structure for each population of interest might need to be obtained, as there will be no ubiquity in structure dictated by, for example, common recombination rates. Finally, it has also been documented that there is considerable individual variation in recombination that could influence LD over-and-above the population-specific and basic biophysical and organizational properties of the human genome (Lynn et al. 2002). Thus, empirically assessing evidence that population-specific factors influence LD strength to a greater degree than fundamental biophysical phenomena associated with chromosome organization and reorganization during meiosis could shed light on aspects of the ubiquity of a haplotype block structure of the human genome. As an example of the confounding of LD patterns induced by recombination rates due to population-specific phenomena, consider recently admixed populations for which the two parental populations undergoing admixture have different allele frequencies. In the early (i.e., pre-equilibrium) generations of admixture, LD will be observed between markers at different loci due to their differential frequencies in the parental populations that may not reflect the degree of recombination in certain genomic regions simply because not enough time has elapsed for this greater recombination to disrupt the LD induced by differential frequencies of alleles across the two populations.

A recent study by Gabriel et al. (2002) suggests that, although the size of the blocks, as well the frequency of the haplotypes observed in the genomic region associated with those blocks, differ across populations, there is an underlying structure to the blocks that appears to be common. Thus, Gabriel et al. (2002) found, among other things, that larger block sizes were found in the same regions of the genome across the populations they studied, although the actual size of these blocks in the relevant genomic regions relative to other genomic regions did vary across the populations. This phenomenon could occur if, for example, recombination were to occur more often in certain re-

[1]Corresponding author.
E-MAIL nschork@ucsd.edu; FAX (858) 822-2113.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1540404.

14:1358–1361 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

gions of the genome, and its effect on LD strength was more pronounced than population-idiosyncratic factors such as the age of the population, population immigration rate, and genetic drift.

Recombination hot spots have been proposed as mechanisms for generating haplotype blocks in many publications (Daly et al. 2001; Cullen et al. 2002; Gabriel et al. 2002; Kauppi et al. 2003). Initial support for this idea was provided by Jeffreys et al. (2001) in a study showing a correlation between recombination hot spots and block boundaries in the major histocompatibility complex class II region. In addition, a recent study by Dawson et al. (2002) demonstrated a correlation between LD strength and recombination for chromosome 22, but their results were only reported for a single chromosome, and haplotype block size was not considered. Therefore, we undertook an evaluation of the relationship between block size and recombination rate using the haplotype block data collected from around the genome by Gabriel et al. (2002) and the genomic recombination rates calculated by Kong et al. (2002).

RESULTS

Figure 1A,B,C offers scatter plots and estimated regression lines between average block size, the size of the largest block observed in a region, and the percentage of DNA sequence in a region falling within blocks and the sex-averaged recombination rate for those regions. Nonparametric Spearman rank correlations between sex-averaged recombination rate and the block parameters were all negative and significant at the 5% type I error level, as shown in Table 1. Similar results were obtained when sex-specific recombination rates were used in the analysis and when Pearson correlations were used (data not shown). Because the correlation measures used make assumptions about the distribution of the test statistics that may not hold for small samples, we also computed *P*-values for our correlations using permutation tests. All *P*-values resulting from the permutation distributions were less than 0.05.

DISCUSSION

Our finding that haplotype block size parameters are negatively correlated with recombination rates is neither obvious nor counterintuitive, but rather represents an empirical finding that could not necessarily be anticipated, as argued extensively by Zhang et al. (2003) and others. For example, a recent simulation study pursued by Wang et al. (2002) demonstrated the importance of population-specific factors in the determination of haplotype block characteristics, and further suggested that it is likely that a combination of evolutionary forces, including mutation, recombination, and demographic history, simultaneously influence haplotype block structure. However, an empirical assessment of these claims requires collection and analysis of the appropriate data.

Our use of genome-wide haplotype block data from populations distinct from that for which recombination frequency data were obtained is compelling in this regard, as we would not

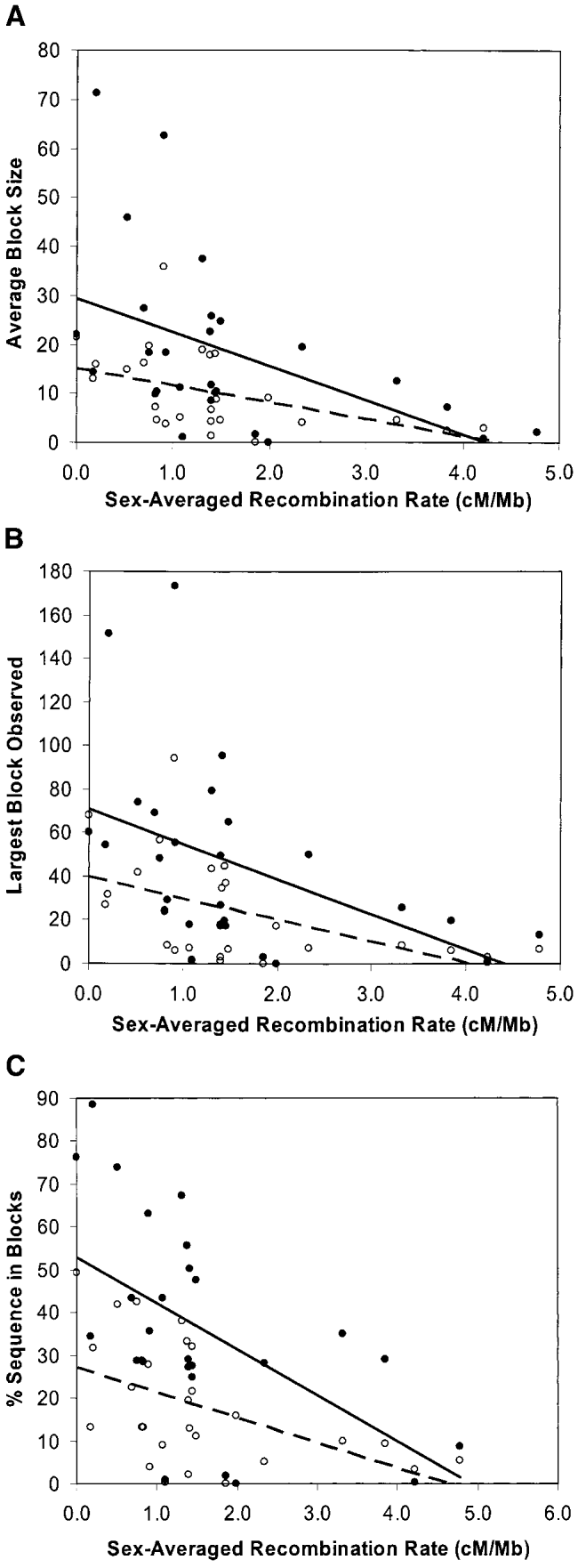


Figure 1 Scatter diagrams depicting the relationship between block structure and recombination rate using the haplotype block data from Gabriel et al. (2002) and the recombination rate data from Kong et al. (2002). (●) Block data obtained from the CEPH families. (○) Block data from the Yoruban samples. The solid and dashed lines are linear regression lines for block structures and recombination rates computed for the CEPH and Yoruban data, respectively. (*A*) The relationship between sex-averaged recombination rate and average block size. (*B*) The relationship between sex-averaged recombination rate and the size of the largest block observed in a region. (*C*) The relationship between sex-averaged recombination rate and percentage of sequence in a region found to be in a block.

Table 1. Spearman Rank Correlations (P-Values) Between Sex-Averaged Recombination Rate and the Three Block Parameters Analyzed

	CEPH	Yoruban
Average block size	−0.55 (<0.01)	−0.59 (<0.01)
Largest block size	−0.59 (<0.01)	−0.56 (<0.01)
Percent sequence	−0.59 (<0.01)	−0.54 (<0.01)

expect to see any correlation between block size and recombination rates if the map generated by Kong et al. (2002) using an Icelandic population was not somehow applicable to the data collected by Gabriel et al. (2002) on the Northern European CEPH and Yoruban populations. Thus, if the recombination rates reported in the Kong et al. (2002) map are not ubiquitous, then our observation of a negative correlation between block size and recombination rate would have to be qualified as an exceptionally improbable and fortuitous phenomenon. Human haplotype block sizes, on the other hand, should show variation across populations because of population-specific factors that influence LD. This suggests that our observation of a negative correlation between block sizes and recombination rates in two different populations is likely due to the ubiquity in recombination rates that influence block sizes. Thus, based on actual data, one might expect similarity in patterns in block structure across different populations due to recombination rate.

Our observed negative correlation invites discussion on one point having to do with the effort and strategies used to identify haplotype blocks within a genomic region. It has been argued that the choice of polymorphism identification strategies can influence estimates of LD and hence block structure (Akey et al. 2003). In addition, there are many different methodologies and algorithms for identifying or labeling haplotype blocks. Finally, it is an open question as to how far one must go to state with certainty that there is no LD—and therefore no blocks—within a certain genomic region. For example, to confidently determine LD patterns, one must pursue an exhaustive search for polymorphism in the regions of interest and then evaluate LD patterns between alleles at all these sites. The Gabriel et al. (2002) data did contain stretches of genome between regions for which block identification was pursued for which blocks were not identified. As these regions were not of focus in their study, they may or may not contain LD blocks. What we can say about our observed finding in this context is that, at least among those regions of the genome where blocks were identified unambiguously with the methodology described by Gabriel et al. (2002), we observe a negative correlation between block size and recombination rate. Given that more and more haplotype block data are being collected by participants in the haplotype map initiative, it is safe to say that there will be plenty of data that can be used to replicate our findings in the future.

Ultimately, although there is tremendous variation in the factors that contribute to LD across populations, as well as great variation in recombination rates between the sexes and among individuals, the genomic regions that undergo frequent recombination are not likely to only manifest this greater recombination in a small subset or specific population of individuals. Thus, any general tendency for greater recombination and variation in a particular genomic region will create pockets of LD and therefore potential haplotype blocks in the population at large. Our observation of a negative correlation between haplotype block parameters from two highly diverse groups and general recombination rates speaks to this phenomenon, and suggests that a potential universality of aspects of a single block-like structure for the human genome might exist. As such, smaller blocks will exist in regions of high recombination for most populations. Confirmation and ultimate use of this structure will undoubtedly emerge as the HapMap Project unfolds.

METHODS

Haplotype Block and Recombination Data

We used the data on LD strength-dictated haplotype block structures discussed by Gabriel et al. (2002), but only considered haplotype block parameters obtained with the CEPH and Yoruban population samples, as they were the most divergent populations studied with respect to block size as reported by Gabriel et al (2002). Microsatellite markers obtained from the Supplemental material associated with the high-resolution genetic map of Kong et al. (2002, www.nature.com/ng Supplemental data table ng917-S13.xls) were chosen to closely flank the regions analyzed by Gabriel et al. (2002) to determine haplotype blocks. The previously determined physical and genetic locations of these microsatellites (Kong et al. 2002) were used for calculations of physical and genetic distance between markers flanking each region and subsequent calculations of sex-averaged and sex-specific recombination rates across each region. For each region for which block data were available, the following measures were determined and ultimately correlated with recombination rate: the average block size, the size of the largest block observed, and the percentage of sequence observed in haplotype blocks. Only those regions for which at least one block was observed in either the CEPH or Yoruban population were included in these analyses. All of this data is available in a Supplemental table.

Statistical Analysis

Simple linear regressions and correlation coefficients (both Pearson's parametric and Spearman's nonparametric) were computed between block structure data and recombination rates. We also used permutation tests to assess the statistical significance of the observed correlations; 1000 data permutations were used for this purpose (Good 1994).

ACKNOWLEDGMENTS

Aspects of this work were supported by NIH NHLBI grants HL54998-01, HL69758-01, and HL64777-03. We thank Stacey Gabriel, Stephen Schaffer, Mark Daly, and David Altshuler (Whitehead/MIT Center for Genome Research) for discussions on the use of their haplotype block data.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Akey, J.M., Zhang, K., Xiong, M., and Jin, L. 2003. The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* **20:** 232–242.

Cullen, M., Perfetto, S.P., Klitz, W., Nelson, G., and Carrington, M. 2002. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am. J. Hum. Genet.* **71:** 759–776.

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

Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., et al. 2002. A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418:** 544–548.

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

Good, P. 1994. *Permutation tests*. Springer, New York.
Jeffreys, A.J., Kauppi, L., and Neumann, R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29:** 217–222.
Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., et al. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29:** 233–237.
Kauppi, L., Sajantila, A., and Jeffreys, A.J. 2003. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.* **12:** 33–40.
Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.
Lewontin, R. 1974. *The genetic basis of evolutionary change*. Columbia University Press, New York.
Lynn, A., Koehler, K.E., Judis, L., Chan, E.R., Cherry, J.P., Schwartz, S., Seftel, A., Hunt, P.A., and Hassold, T.J. 2002. Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science* **296:** 2222–2225.
Olivier, M., Bustos, V.I., Levy, M.R., Smick, G.A., Moreno, I., Bushard, J.M., Almendras, A.A., Sheppard, K., Zierten, D.L., Aggarwal, A., et al. 2001. Complex high-resolution linkage disequilibrium and haplotype patterns of single-nucleotide polymorphisms in 2.5 Mb of sequence on human chromosome 21. *Genomics* **78:** 64–72.
Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P., et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294:** 1719–1723.
Stumpf, M.P. 2002. Haplotype diversity and the block structure of linkage disequilibrium. *Trends Genet.* **18:** 226–228.
Wang, N., Akey, J.M., Zhang, K., Chakraborty, R., and Jin, L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71:** 1227–1234.
Zhang, K., Akey, J.M., Wang, N., Xiong, M., Chakraborty, R., and Jin, L. 2003. Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: An act of genetic drift. *Hum. Genet.* **113:** 51–59.

WEB SITE REFERENCES

http://www.hapmap.org; The International Haplotype Map Project.
http://www.nature.com/ng/; *Nature Genetics*.

Received May 14, 2003; accepted in revised form April 2, 2004.