Human Molecular Genetics, 2003, Vol. 12, No. 7 ***771–776***
DOI: 10.1093/hmg/ddg088

Linkage disequilibrium patterns of the human genome across populations

Sagiv Shifman[1,2], Jane Kuypers[3], Mark Kokoris[3], Benjamin Yakir[4] and Ariel Darvasi[1,2,*]

[1]The Life Sciences Institute, The Hebrew University of Jerusalem, Jerusalem 91904, Israel , [2]IDgene Pharmaceuticals Ltd, Jerusalem 91344, Israel , [3]QIAGEN Genomics Inc., 1725 220th St SE, Bothell, WA 98021, USA and [4]Department of Statistics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Received December 9, 2002; Revised January 30, 2003; Accepted February 5, 2003

We studied the patterns of linkage disequilibrium (LD) in the human genome among three populations: African Americans, Caucasians and Ashkenazi Jews. These three populations represent admixed, outbred and isolated populations, respectively. The study examined defined chromosomal regions across the whole genome. We found that SNP allele frequencies are highly correlated between Ashkenazi Jews and Caucasians and somewhat distinct in African Americans. In addition, Ashkenazi Jews have a modest increase in LD compared with Caucasians, and both have greater LD than African Americans. The three populations differed more significantly with regard to haplotype heterogeneity. We found, as expected, that Ashkenazi Jews display the greatest extent of homogeneity and African Americans the greatest extent of heterogeneity. We found that most of the variance in LD can be attributed to the difference between regions and markers rather than to that between different population types. The average recombination rates estimated by low-resolution genetic maps can only explain a small fraction of the variance between regions. We found that LD (in terms of r^2) decreases as a function of distance even within the so-called 'haplotype blocks'. This has significant consequences when using LD mapping for the genetic dissection of complex traits, as higher density SNP maps will be required to scan the genome.

INTRODUCTION

Linkage disequilibrium (LD)—the non-random association between alleles of different loci—may be extremely important for the dissection of complex traits. Since the probability of association studies in testing the functional polymorphism may be low, it is essential to rely on other polymorphisms, which are in LD with the functional polymorphism. Although these polymorphisms are associated with the trait, they do not directly influence it. High LD is nevertheless a two-edged sword: while on the one hand it increases the chances of identifying an associated region, it also interferes with fine mapping and the identification of the susceptibility gene. Great debate exists regarding the extent of LD in the human genome and among various populations. Recent studies have suggested that the human genome is organized in blocks of haplotypes with high LD, separated by regions of low LD (1,2). The size of the blocks that show no evidence of recombination was found to be smaller in African than in European and Asian populations (3).

The level and pattern of LD is influenced by many factors on several levels of comparison. LD varies across populations and genome regions and between pairs of markers in close proximity (4). Some of the factors which generate LD variance, e.g. genetic drift, admixture and inbreeding, are population specific. Other factors, such as recombination rate, gene conversion and natural selection, are specific to the genomic region (5). Pairs of markers may have different histories, such as the age of the mutation, and are also influenced by different genomic processes.

The main force which breaks down LD is recombination. Recent evidence has accumulated indicating that recombination events are not uniformly distributed along the human genome but are localized in hot spots (6). If recombination acts in such a binary way, LD will break down according to hot spot density and recombination rates. Regions with the same average recombination rate can demonstrate highly different levels of LD if the density of their recombination hot spots differs widely.

Three types of populations were suggested to have an increased LD: (i) admixture of two populations with different allele frequencies (7); (ii) small stable populations (8); and (iii) isolated populations founded by a small number of people, which have since grown rapidly (9,10). Using computer simulations, Kruglyak (11) has demonstrated that, unless the

*To whom correspondence should be addressed. Tel: +972 26595600; Fax: +972 26595601; Email: arield@cc.huji.ac.il

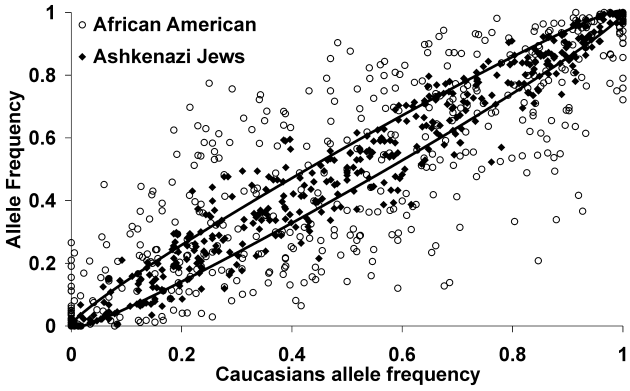

Figure 1. Comparison of allele frequencies across populations. Allele frequencies in samples of African Americans and Ashkenazi Jews plotted against allele frequencies in Caucasians. The solid lines show 95% confidence interval for allele frequency estimation in Caucasians.

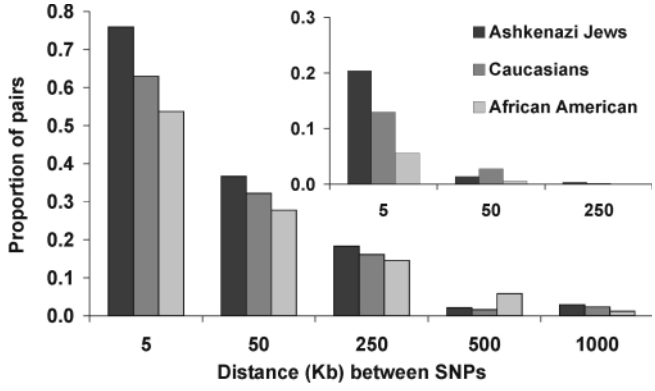

Figure 2. The proportion of SNP pairs showing no evidence of recombination ($D' = 1.0$) plotted for different distance intervals between SNPs. Inset: the proportion of SNP pairs displaying only two out of the four possible haplotypes ($r^2 = 1.0$).

haplotypes were introduced by a small number of founders, LD should be similar in outbred and isolated populations. The number of founders carrying each haplotype is a function of the founder-population size and haplotype frequencies.

Here we present a systematic estimation of LD in the human genome in three populations: African Americans, an admixed population; Caucasians, an outbred population of European ancestry; and Ashkenazi Jews, an isolated population.

RESULTS

We studied the extent of LD by choosing two regions of 1 Mb on each of the human chromosomes. In each region we chose a constant distribution of 16 SNPs, providing diversely spaced pairs of SNPs. Out of the 722 successfully designed assays, 602 (83.4%) were successfully genotyped, 384 (63.8%) of the total being polymorphic (minor allele frequency >5%) in all populations. We genotyped 90 individuals from each population and estimated the haplotypes frequencies.

Allele frequencies similarity

We observed very similar allele frequencies in the Ashkenazi and Caucasian samples ($R^2 = 0.96$), both of which differed from the African American sample (Fig. 1, $R^2 = 0.71$). Among the non-polymorphic SNPs in the Caucasian and Ashkenazi samples, 50% were polymorphic in the African American sample. Only 18% of the non-polymorphic SNPs in the African American sample were polymorphic in the Caucasian and Ashkenazi samples.

Measures and level of LD

We calculated two measures of LD (D' and r) for each of the 1356 pairs of SNPs. Measuring LD with r or r^2 possesses several advantages over D'. While D' is biased upward in small sample sizes and for low allele frequencies, r exhibits more reliable sampling properties. When testing a marker in LD with the functional polymorphism, sample size should be increased in proportion to $1/r^2$ in order to obtain the same power as directly testing the functional polymorphism (12). r^2 is therefore more appropriate when association studies are of interest. For pairs of biallelic markers, D' will equal 1.0 when one or two out of the four possible haplotypes are missing from the population, r equaling 1.0 when there are only two haplotypes. When no recombination has occurred between two markers, D' will equal 1.0 (in the absence of mutation or genotyping error), while r will be dependent on both markers' allele frequencies. D' measure is therefore used to model recombination rates and r or r^2 to model association power.

The proportion of SNP pairs showing no evidence of recombination ($D' = 1.0$) decreases dramatically with distance (Fig. 2). SNPs separated by less than 5 kb showed no evidence of recombination for 76, 63 and 54% of the pairs in the Ashkenazi, Caucasian and African American samples respectively. At larger distances (40–80 kb), only ~14% of the pairs showed no evidence of recombination.

While the average LD between SNP pairs declined at a similar rate in the Ashkenazi and Caucasian samples, it declined more rapidly in the African American samples (Fig. 3). Although on the average r^2 is only modestly elevated in the Ashkenazi samples compared with the Caucasian, the proportion of pairs where $r^2 = 1.0$ (SNP pairs with only two haplotypes) differs in the three populations (Fig. 2, inset). We found that 20, 13 and 6% of SNP pairs separated by less than 5 kb in the Ashkenazi, Caucasian and African American, respectively, were in absolute LD ($r^2 = 1.0$). Almost none of the SNP pairs separated by more then 10 kb were in such absolute LD.

We calculated the average r^2 for SNP pairs which demonstrated low evidence of recombination and therefore lie in the same haplotype block (3). Since D' values are biased upward when a small sample or rare alleles are examined, we used only SNP pairs for which the observed D' value is above 0.98 and for which the expected D' value is significantly above 0.9 in all three populations. This constitutes a more stringent criterion to define a haplotype block than the one used by Gabriel *et al*. (3). These selected pairs were chosen to represent very low levels of between-SNP recombination. Surprisingly, the average r^2 significantly declines within a haplotype block as a function of distance between the SNPs (in Ashkenazi Jews and Caucasians, r^2 declines with distance from 0.7–0.8 to

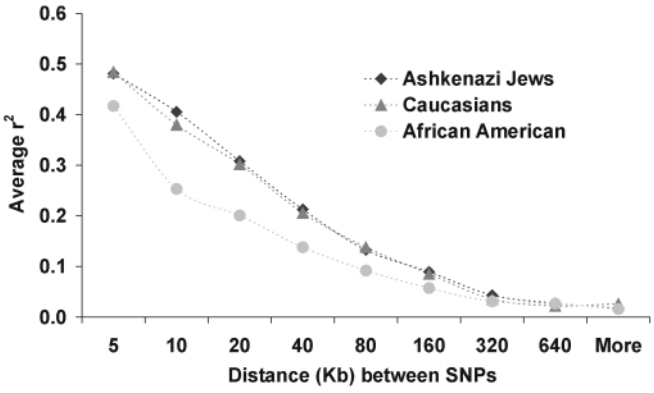

Figure 3. Average LD calculated by r^2 plotted for different distance intervals between SNPs in the different populations.

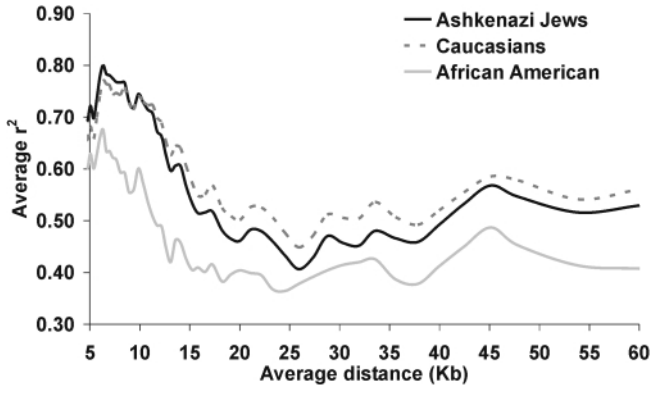

Figure 4. The effect of distance on average LD calculated by r^2 for SNP pairs within 'haplotype blocks'. r^2 is presented by a moving average of 20 SNP pairs.

Table 1. Proportion of LD variance explained by physical and genetic distance and GC content

	Physical distance	Genetic distance		GC content and genetic distance
		Marshfield	deCODE	
Ashkenazi Jews	28.4%***	25.7%^{NS}	29.8%***	30.5%**
Caucasians	30.1%***	27.3%^{NS}	32.2% ***	32.6%*
African American	19.5%***	15.8%^{NS}	20.6% ***	20.7%^{NS}

NS, not significant; *$P < 0.01$; **$P < 0.001$; ***$P < 0.0001$. The significance for 'Genetic distance' refers to the addition over 'Physical distance'. The significance of 'GC content and genetic distance' refers to the addition over 'Genetic distance' alone.

0.4–0.6, and in African Americans from 0.6–0.7 to 0.4–0.5; see Fig. 4).

We calculated the correlation between LD, measured by *r*, for SNP pairs across the three populations. We found that LD (magnitude and sign) is highly correlated across populations: the LD variation of the Caucasian sample explains 63 and 82% of the LD variance in the African American and Ashkenazi samples, respectively.

Recombination rate, GC content and LD

Recombination rates should be a good predictor of LD level in different regions. We tested the correlation between LD results (D') and the genetic distance, the latter obtained from the regions' average recombination rates (cM/Mb), estimated by the Marshfield genetic map and the recently published improved estimations by deCODE Genetics (13). In addition, we tested the correlation between GC content and LD (Table 1). We found that the recombination rate estimated by deCODE's genetic maps provides a significant additional predictor of LD beyond the physical map, while Marshfield's recombination rates were not significantly correlated with our LD results. The GC content was also found to be a significant additional predictor of LD, proving to be an important predictor in the Ashkenazi sample, a lower level predictor in Caucasians, and an insignificant LD predictor in African Americans. Although recombination rates are highly significant correlated with LD, they provide a small fraction of additional variance explained (1.4–3.1%). The genetic distance was not highly correlated with the LD level of SNPs pairs separated by more than 250 kb across all populations or in SNP pairs separated by less then 5 kb in the Caucasian and Ashkenazi samples (Fig. 5). In the African American sample, in which LD drops more rapidly, the correlation between LD level and genetic distance is not pronounced, even for distances of more then 50 kb. The highest correlation between LD levels and genetic distance is observed for SNPs separated by 5–50 kb ($R^2 = 0.13$).

The extent of useful LD

We estimated the effective population size (N_e) for each region by the commonly accepted model, which assumes a random mating population in recombination-drift equilibrium (14). The mean r^2 between SNP pairs is given by $1/(1 + 4N_e\theta)$, where θ is the recombination rate between SNP pairs. We estimated the expected distance, which corresponds on the average to $r^2 = 1/3$, by using the above model (Fig. 6). It has been suggested that a maximal threefold increase in sample size (i.e. r^2 above 1/3) will be essential for association studies. Under this assumption and a genome average recombination ratio of 1 Mb = 1 cM, the extent of LD useful for association studies ranges from 197, 153 and 138 kb in the Ashkenazi, African American and Caucasian samples, respectively, in the highest LD level region, to less than 1 kb in the region with the lowest LD level. Half of the regions show useful LD beyond 6 kb in the African Americans sample, and beyond 11 kb in the Ashkenazi and Caucasians samples. In order to present the average extent of LD with less variance between regions, we divided the

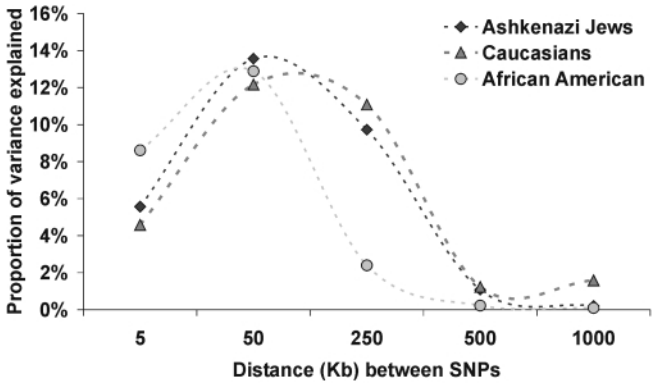

Figure 5. The proportion of variance of LD explained by the genetic distance, calculated separately for different distance intervals between SNP pairs.

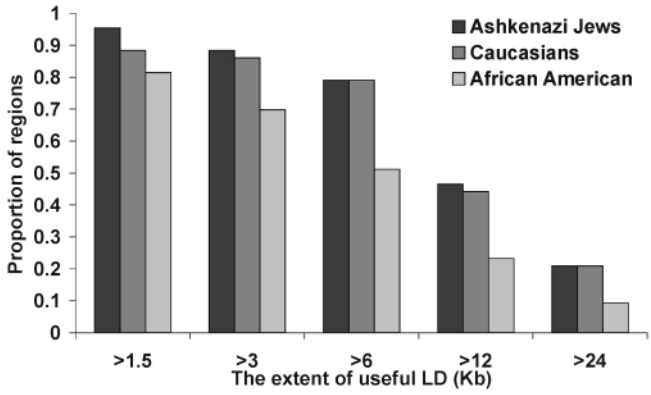

Figure 6. The cumulative proportions of regions for which useful LD was observed, presented as a function of the distance between SNP pairs. The average extent of useful LD was estimated for each region by applying the LD decay model and calculating the expected distance for which the average LD (r^2) drops below 1/3.

regions into two even groups of high and low LD according to the estimated LD level averaged across the three populations, referred below as the 'high LD regions' and the 'low LD regions'. The distribution of LD (r^2) is plotted separately for the high and low LD regions in Fig. 7. In the high LD regions, the average extent of LD useful for association is 24.4, 23.2 and 12.2 kb for the Ashkenazi, Caucasian and African American populations, respectively. In the low LD regions it extends to 6.7, 6.5 and 2.5 kb for the Ashkenazi, Caucasian and African American populations, respectively.

DISCUSSION

We studied LD between common polymorphisms among African Americans, Caucasians and Ashkenazi Jews. We found a high similarity of allele frequencies and LD level between Caucasians and Ashkenazi Jews. The African American sample showed a higher diversity and lower levels of LD. We found a modest increase in LD level in Ashkenazi Jews compared with the level in the Caucasian group. In a previous study of a specific region of chromosome X, we reported a much greater difference in LD level between Ashkenazi Jews and Caucasians for SNPs separated by large distance (10). However, this was partially a consequence of the inappropriate mode of analysis (15). The high similarity of LD patterns between Ashkenazi Jews and Caucasians suggests that the same historical events generated most of the extended LD in these two populations. Recently, Gabriel *et al.* (3) found a high degree of concordance between Asians and Europeans in recombination evidence between SNP pairs. Africans were found to be substantially less concordant with Europeans and Asians. The high similarity in patterns of LD across non-African populations may suggest that the historical event—probably a severe bottleneck which generated the extended LD in non-African populations—occurred after the 'Out of Africa' event of modern humans (3).

The African American population represents an admixed population with genetic contributions from both African and European ancestors. It has been suggested that the African American population will be useful for LD mapping, since populations of recently mixed ethnic groups display linkage disequilibrium over long intervals (16). We did not, however, observe high LD between SNPs separated by long distance in this population. On the contrary, similarly to findings presented by Gabriel *et al.* (3), where LD patterns of Africans and African Americans were found to be similar, we found levels of LD to be low in this population. The strong LD in admixed populations should be pronounced only in markers which possess a large allele frequency difference between the parental populations contributing to the admixed population. In addition it has been demonstrated that the degree of Caucasian admixture in the African American population differs between geographic locations (17). Our samples, contributed from an anonymous panel cannot, therefore, completely dismiss the value of specific African American populations as an admixed population for mapping complex trait genes.

SNPs common in all populations emerged from old mutations and can be used for the study of early human evolution. In regions with very low recombination rates, some of the haplotype combinations for these SNPs are relatively rare or absent in a particular population due to genetic drift and recent historical events. Some of the haplotypes present in the Caucasian sample are absent from the Ashkenazi sample. Most of these haplotypes are of SNPs separated by a short distance and are relatively rare in the Caucasian sample. This data is consistent with historical studies suggesting that the Ashkenazi population descended from a small number of founders and then rapidly expanded 700–1000 years ago. The Ashkenazi population-founder bottleneck eliminated many rare haplotypes and alleles, causing a reduction in the population's genetic heterogeneity and a modest increase in LD. Haplotypes of widely spaced SNPs are reconstructed by recombination and thus only haplotypes of tightly spaced SNPs are expected to be absent. The effect of allelic and non-allelic heterogeneity on the power to detect complex traits genes and the advantage of isolated populations in that context has been discussed elsewhere (10,18).

The correlation between genetic distance and LD is low for SNPs separated by very short distance or long distance. SNPs separated by long distance are for the most part in linkage equilibrium and therefore the measurements of LD in a sample are not correlated with genetic distance. Since recombination is

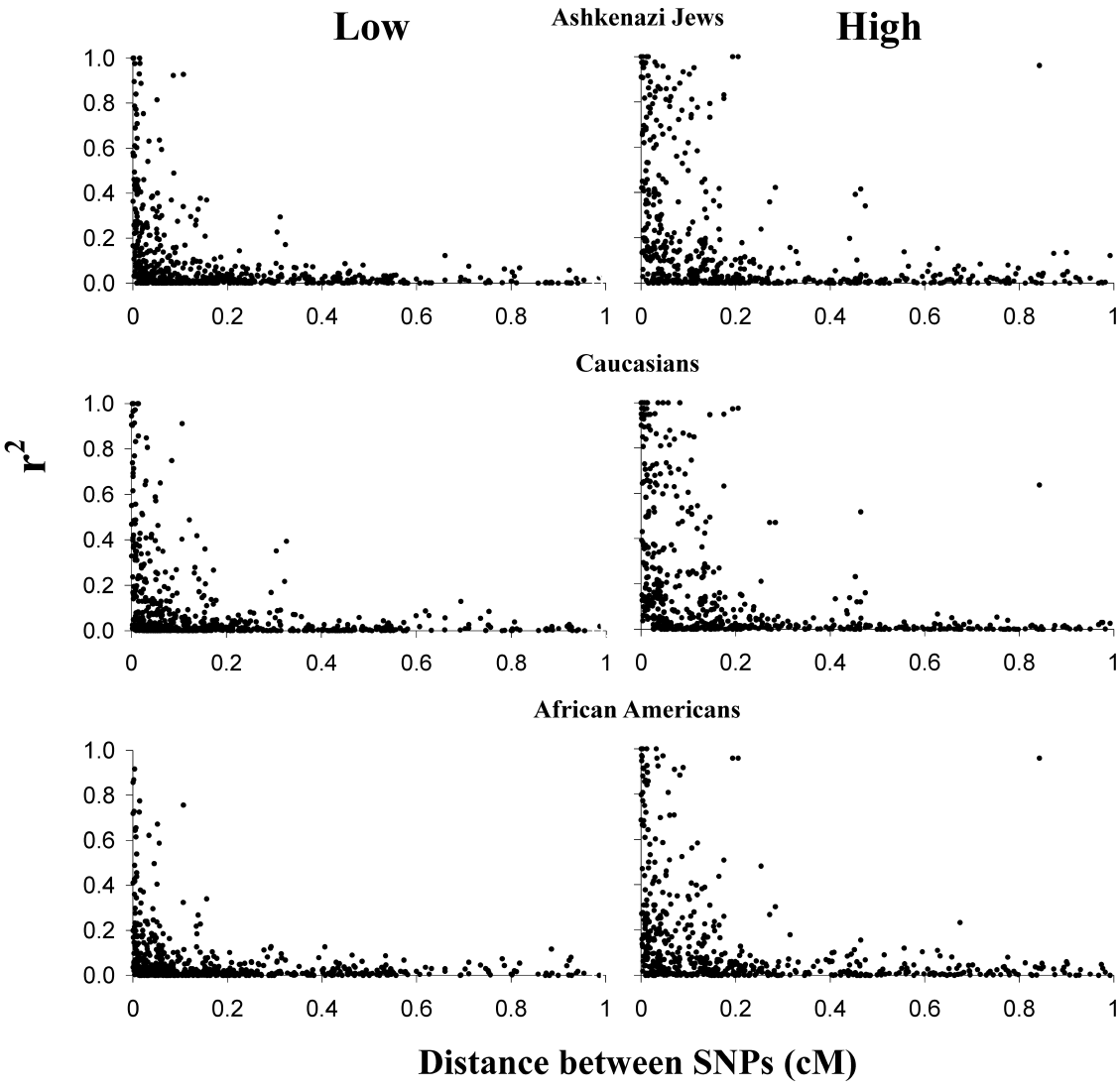

Figure 7. LD distribution presented by r^2 as a function of distance (in cM) for Ashkenazi Jews, Caucasians, and African Americans. The regions were divided into two groups of high and low LD regions according to the average extent of useful LD across the three populations. Regions where the useful level of LD extends beyond an average distance of 8.5 kb were plotted together as high LD regions and those with useful LD below 8.5 kb as low LD regions.

rare between SNPs separated by short distance, the LD level between these SNPs is also relatively uninfluenced by the genetic distance. Following this argument, it has been suggested that LD (measured by r^2) does not substantially decline with distance within haplotype blocks (3). We found that r^2 decreases as a function of distance, inclusive of pairs of SNPs which can be defined as belonging to the same haplotype block. Decrease in r^2 within blocks between 0 and 40 kb can also be observed by a close examination of Gabriel *et al.*'s (3) data (Fig. 2C). A gradual decay of LD within blocks has also been observed in the HLA region, suggesting that crossovers are not entirely restricted to hot spots (19). Knowledge of the boundaries of haplotype blocks alone may consequently be insufficient for association mapping. While LD decay within blocks offers further resolution for fine mapping, it also requires higher SNP density to scan the genome.

We found that LD varies extensively between different regions and between SNP pairs in a non-predictable way. Although we divided the regions into high and low LD areas, in practice the distribution is continuous. For association studies to be feasible, with an optimal distribution of markers, the level of LD should be estimated empirically for each region. Since the extensive variance in LD seems to be highly correlated between populations, construction of an LD map of the human genome should provide a useful tool for the design of cross-population association studies. Currently, low-resolution genetic maps may be used to predict LD levels to a very limited degree.

The extent of useful LD in the low LD regions approximates the distance estimated by simulations (11). Few explanations can account for the high LD level observed in some regions and the high variability between regions, and a simple demographic model seems insufficient. If recombination is localized in hot spots in some regions, the level of LD in those regions will be influenced by hot spot density and recombination rate in each hot spot, but only partially by the average recombination rate of the region. In regions in which recombination is uniformly

distributed, we expect LD to be low, as predicted by simulation based on the average recombination rate. This may explain the high variability we observed in cross-regional LD levels even when an average recombination rate is introduced.

MATERIAL AND METHODS

Samples

The study consisted of 270 samples, 90 individuals from each of the three populations. The Ashkenazi sample consisted of individuals from Israel, both of whose parents were born in Eastern Europe. The African American and Caucasian samples were obtained from the NIGMS Human Genetic Cell Repository (HD100AA, HD100CAU, Corriel Cell Repositories, Camden, NJ, USA).

Marker selection

In each chromosome we selected two regions of 1 Mb. Regions were selected to include gene-poor and gene-rich areas at different locations on the chromosome. SNPs were selected from the NCBI SNP database according to a constant distribution in each selected region. We selected 16 SNPs with a distribution closest to the following distances (kb) between the SNPs: 300, 150, 5, 5, 5, 15, 15, 15, 15, 15, 5, 5, 5, 150 and 300.

Genotyping method

Genotyping was carried out using the Masscode™ system (QIAGEN Genomics) as previously described (20). The Masscode™ SNP discrimination assay is a two-step PCR-based process. The primary reaction, which requires only standard PCR primers, amplifies a 100–300 bp region spanning the SNP site. The second PCR reaction is a competitive allele-specific PCR which uses two, non-modified and unpurified allele-discrimination primers paired with differentially tagged universal Masscode™ primers to generate the genotype result. The photocleavable Masscode™ tags are uniquely associated with the SNP site via oligo tails synthesized at the 5′ end of each allele-specific primer. Following a DNA cleanup and photolysis step to cleave the tags from the amplicon, the processed samples are acquired using a single quadrupole mass spectrometer. Masscode™ data is analyzed using a custom software package called DataGen™, which determines the genotype based on the relative proportions of the paired allele tags.

Statistical analysis

We estimated the haplotype frequencies using the expectation maximizations algorithm. We measured LD between SNP pairs using the absolute value of Lewontin's D' ($|D'|$) and Pearson correlation (r). The effective population size parameter (N_e) was estimated by a non-linear regression. The recombination rate (θ) was estimated by the physical distance and the regional average recombination rate.

SNP pairs were classified as pairs with low evidence of recombination if the estimated D' value was above 0.98 and was significantly higher than 0.9 in all populations. P-values were calculated via simulation. Since we tested many SNP pairs, we used the false discovery rate linear step-up procedure (21) in order to control for a total error rate below 0.05.

REFERENCES

1. Goldstein, D.B. (2001) Islands of linkage disequilibrium. *Nat. Genet.*, **29**, 109–111.
2. Daly, M.J., Rioux, J.D., Schaffner, S.E., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.
3. 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.
4. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
5. Ardlie, K.G., Kruglyak, L. and Seielstad, M. (2002) Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.*, **3**, 299–309.
6. Stumpf, M.P.H. (2002) Haplotype diversity and the block structure of linkage disequilibrium. *Trends Genet.*, **18**, 226–228.
7. Smith, M.W., Lautenberger, J.A., Shin, H.D., Chretien, J.P., Shrestha, S., Gilbert, D.A. and O'Brien, S.J. (2001) Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. *Am. J. Hum. Genet.*, **69**, 1080–1094.
8. Terwilliger, J.D., Zollner, S., Laan, M. and Paabo, S. (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum. Hered.*, **48**, 138–154.
9. Sheffield, V.C., Stone, E.M. and Carmi, R. (1998) Use of isolated inbred human populations for identification of disease genes. *Trends Genet.*, **14**, 391–396.
10. Shifman, S. and Darvasi, A. (2001) The value of isolated populations. *Nat. Genet.*, **28**, 309–310.
11. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
12. Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.*, **69**, 1–14.
13. 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.
14. Hill, W.G. and Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, **38**, 226–231.
15. Maniatis, N., Collins, A., Xu, C.F., McCarthy, L.C., Hewett, D.R., Tapper, W., Ennis, S., Ke, X. and Morton, N.E. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2228–2233.
16. Stephens, J.C., Briscoe, D. and O'Brien, S.J. (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am. J. Hum. Genet.*, **55**, 809–824.
17. Destro-Bisol, G., Maviglia, R., Caglia, A., Boschi, I., Spedini, G., Pascali, V., Clark, A. and Tishkoff, S. (1999) Estimating European admixture in African Americans by using microsatellites and a microsatellite haplotype (CD4/Alu). *Hum. Genet.*, **104**, 149–157.
18. Wright, A.F., Carothers, A.D. and Pirastu, M. (1999) Population choice in mapping genes for complex diseases. *Nat. Genet.*, **23**, 397–404.
19. 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.
20. Kokoris, M., Dix, K., Moynihan, K., Mathis, J., Erwin, B., Grass, P., Hines, B. and Duesterhoeft, A. (2000) High-throughput SNP genotyping with the Masscode system. *Mol. Diagn.*, **5**, 329–340.
21. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodol.*, **57**, 289–300.