A Scan for Linkage Disequilibrium Across the Human Genome

Gavin A. Huttley,* Michael W. Smith,‡ Mary Carrington‡ and Stephen J. O'Brien*

**Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland 21702 and ‡Intramural Research Support Program, SAIC Frederick, Frederick, Maryland 21702*

Manuscript received September 5, 1998
Accepted for publication April 27, 1999

ABSTRACT

Linkage disequilibrium (LD), the tendency for alleles of linked loci to co-occur nonrandomly on chromosomal haplotypes, is an increasingly useful phenomenon for (1) revealing historic perturbation of populations including founder effects, admixture, or incomplete selective sweeps; (2) estimating elapsed time since such events based on time-dependent decay of LD; and (3) disease and phenotype mapping, particularly for traits not amenable to traditional pedigree analysis. Because few descriptions of LD for most regions of the human genome exist, we searched the human genome for the amount and extent of LD among 5048 autosomal short tandem repeat polymorphism (STRP) loci ascertained as specific haplotypes in the European CEPH mapping families. Evidence is presented indicating that ~4% of STRP loci separated by <4.0 cM are in LD. The fraction of locus pairs within these intervals that display small Fisher's exact test (FET) probabilities is directly proportional to the inverse of recombination distance between them (1/cM). The distribution of LD is nonuniform on a chromosomal scale and in a marker density-independent fashion, with chromosomes 2, 15, and 18 being significantly different from the genome average. Furthermore, a stepwise (locus-by-locus) 5-cM sliding-window analysis across 22 autosomes revealed nine genomic regions (2.2–6.4 cM), where the frequency of small FET probabilities among loci was greater than or equal to that presented by the HLA on chromosome 6, a region known to have extensive LD. Although the spatial heterogeneity of LD we detect in Europeans is consistent with the operation of natural selection, absence of a formal test for such genomic scale data prevents eliminating neutral processes as the evolutionary origin of the LD.

LINKAGE disequilibrium (LD) occurs in populations as a consequence of mutation, random genetic drift, selection of single or linked alleles, and population admixture (see Hartl and Clark 1990). Although traditional interest in LD was in recapitulation of historic demographic and selective events, more recently the signals of LD association have been employed in identifying hereditary disease genes in populations as an adjunct to traditional pedigree mapping analysis (Hastbacka *et al.* 1992; Briscoe *et al.* 1994; Stephens *et al.* 1994; Ewens and Spielman 1995; Jorde 1995; Kaplan *et al.* 1995).

Mapping association studies explicitly depend upon the persistence of LD, which decays at a rate proportional to the recombination fraction between the two loci in LD and the number of generations, *G*, since the establishment of LD (Ewens 1979; Hartl and Clark 1990). The dependence of decay in LD on the recombination fraction and *G* have also been exploited to estimate the time elapsed since the initial event that established LD in the ancestral population (Kaplan *et al.* 1994; Tishkoff *et al.* 1996; Stephens *et al.* 1998).

Corresponding author: Gavin A. Huttley, Human Genetics Group, John Curtin School of Medical Research, The Australian National University, Canberra ACT, 0200, Australia.
E-mail: gavin.huttley@anu.edu.au

Different evolutionary origins of LD may produce different genomic patterns among selectively neutral loci. For instance, genetic drift will cause regions of LD randomly distributed across the entire genome. The number of genes in LD within a region, and thus the physical extent of LD, will depend on effective population size and the local recombination rate. Genetic drift may contribute to admixture LD, which arises when genetically differentiated populations interbreed. Admixed LD will exist between those loci that genetically distinguish, by virtue of allele frequency differences, the ancestral populations. Where the genetic differentiation arose from the operation of genetic drift in each ancestral population, the resulting LD also occurs randomly across the genome and potentially over substantial physical distances for a small number of generations (Chakraborty and Weiss 1988; Briscoe *et al.* 1994; Stephens *et al.* 1994).

In contrast to the unbiased distribution of LD from drift, the operation of mutation or natural selection may affect the genomic pattern of LD in a nonuniform way. While genomic regions with high mutation rates at neutral loci are expected to exhibit less LD (Slatkin 1994), natural selection can produce very localized concentrations of LD. For example, if a rare genetic variant at a locus becomes the subject of directional positive selection, then alleles at linked neutral markers will

also increase in frequency, resulting in LD among the hitchhiking loci. Such directional positive selection is frequently referred to as a selective sweep. Completion of a selective sweep is fixation of both the favored variant and its flanking genetic background, thus eliminating LD in the region. Epistatic selection among linked genes may also lead to linkage disequilibrium between flanking neutral loci, depending on the age of the interaction (Lewontin and Kojima 1960; Lewontin 1964; Wiehe and Slatkin 1998). In regions of epistatic interactions, LD among neutral markers may persist if episodic fluctuations in selection are common. Thus, differential genomic patterns of LD among neutral loci are expected under various evolutionary scenarios.

For human population analysis, studies of LD have been limited by a paucity of available human markers and knowledge of their genotypic phase. Recent efforts to assess the background pattern of LD in humans have employed a small number of markers localized to specific genomic regions (Peterson *et al.* 1995; Laan and Paabo 1997). Here we analyze the extent of LD that occurs among 5048 short tandem repeat polymorphism (STRP) loci distributed over all autosomes resolved by the GÉNÉTHON gene mapping project using the European Utah and Amish Centre d'Etude du Polymorphisme Humain (CEPH) families (Weissenbach *et al.* 1992; Gyapay *et al.* 1994; Dib *et al.* 1996). The study had two objectives: (1) assess the relationship between LD and recombination fraction (centimorgans) in the human genome; and (2) inspect the entire human genome to identify and characterize in strength and centimorgan length, regions of remarkable LD. The first analysis involved Fisher's exact tests (FETs) for independence of all locus pairs separated by ≤30 cM; the second involved a statistical procedure for quantifying clustered LD that corrects for marker density (see materials and methods). The results identify considerable LD, a striking inverse proportionality between LD and recombination distance (centimorgans), and 10 chromosomal regions that display substantially elevated LD in the human genome.

MATERIALS AND METHODS

Data and haplotype determination: Genotype data for 5048 STRP loci resolved by the GÉNÉTHON gene mapping project using the European Utah and Amish CEPH families 1331, 1332, 1347, 1362, 1413, 1416, and 884 (Weissenbach *et al.* 1992; Gyapay *et al.* 1994; Dib *et al.* 1996) were obtained from http://www.genethon.fr/genethon_en.html/. Each family consists of three generations: four grandparents, two parents, and from 9 to 15 grandchildren. Although genotype data are available for an additional Venezuelan family, these data were excluded due to evidence of population admixture (Begovich *et al.* 1992; Moonsamy *et al.* 1997). It was suggested that the Amish are differentiated from the Utah families (McLellan *et al.* 1984). However, an exact test for population substructure (Raymond and Rousset 1995), performed using 50 STR loci from chromosomes 1 and 2 separated by ≥5 cm, failed to reject the null hypothesis that these families belong to the same population ($P = 0.76$).

Pedigree information was used to determine phase of the grandparental chromosomes for all markers in the families, producing a total of 54 independent chromosomal haplotypes. Grandparental origin of alleles in parents was readily discernible from grandparent(s)-parent combinations in ~94% of cases. For the small proportion of loci (<0.25%) that could not be resolved anywhere in a pedigree, the allele was classified as missing data. This occurred when all parent-offspring combinations were identical heterozygotes. In the remaining unresolved cases (<6%), although the grandparent(s)-parent combinations were identical heterozygotes, some parent-grandchild combinations were resolvable. To determine the grandparental phase for a single such unresolved locus within a family, one informative closely linked locus on each side of the locus of interest was identified. We define informative in this context as a locus that is segregating at least two alleles in the pedigree and for which phase was unambiguous. The subset of haplotype combinations in the pedigree that contained the grandparental alleles at the informative loci was determined. In <0.25% of cases, more than one allele was present at the phase unresolved locus in the haplotype subset. In this rare instance, the allele present on >60% of haplotypes was selected; otherwise a missing data allele was assigned (<0.13% of cases).

Previous work has shown that the major histocompatibility complex (MHC) exhibits evidence for extensive LD among STR loci and can thus serve as a reference for the rest of the genome (Carrington *et al.* 1998). Because the GÉNÉTHON map contains only three markers in this region (D6S291, D6S273, and D6S265), we genotyped the same CEPH families for six non-GÉNÉTHON markers (MogCA, MIB, DQCAR, G51152, TAP1CA, and RING3CA) located in the MHC (Figure 1; Foissac *et al.* 1997; Martin *et al.* 1998). GÉNÉTHON sex-averaged genetic map distances were utilized as our marker order reference for most analyses (Dib *et al.* 1996). The genetic map for chromosome 6 was modified for marker order around the MHC using DNA sequence data and YAC contig and radiation hybrid (RH) data from the Whitehead Institute (release 11, www-genome.wi.mit.edu/).

Statistics: Statistical significance determined by FETs, rather than association statistics, was used to measure LD. Historically, measuring LD has been performed for biallelic loci using the coefficient of LD D, or derivatives such as D' or r^2 (Lewontin 1964; Hill and Robertson 1968), with large values of the statistics interpreted as representing significant LD. Multiallelic formulations of these statistics are also available (see, for example, Hedrick 1987; Klitz *et al.* 1995). Yet, as discussed by Slatkin (1994), small values of D may also be associated with significant LD. Furthermore, interpreting measures of association is often problematic (Press *et al.* 1992, p. 631), and under some circumstances the distributions of the statistics can differ substantially from that assumed (Hudson 1985; Hedrick 1987). In contrast, the principal limitation of probabilities from significance tests for measuring LD is their sensitivity to the marginals (row and column sums) of the pairwise tables, and hence sample size (Bennett and Hsu 1960). Two issues arise from a consideration of the influence of sample size. First, while a low probability may be taken as evidence of LD, a high probability cannot be interpreted as evidence of linkage equilibrium. Second, the sensitivity to sample size suggests that loci with large numbers of alleles, and thus high heterozygosity, will have a reduced power to detect LD because of the increased likelihood of unique alleles at such loci. Contradicting this are the simulation results of Slatkin (1994), who showed that, for sample sizes of 100 chromosomes, loci with a large number (eight) of alleles had

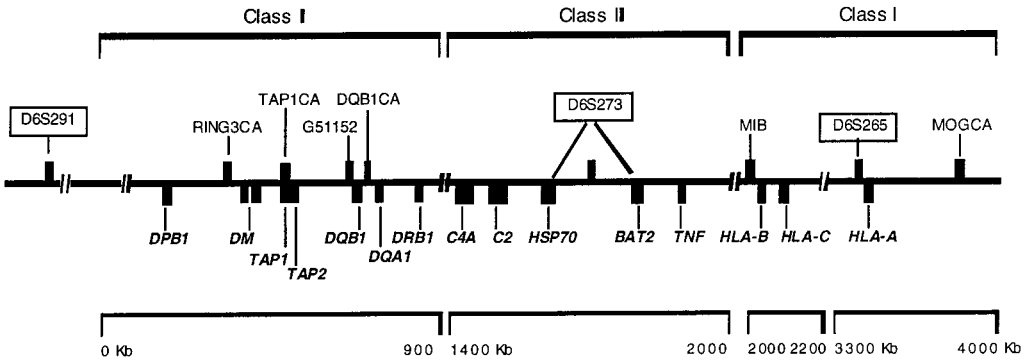

Figure 1.—Map of the HLA complex. STRP loci are shown above, and coding genes below, the map with GÉNÉTHON markers in boxes.

substantially greater power to detect significant LD than biallelic loci. Interestingly, the relationship between power to detect LD and number of alleles approaches a plateau after triallelic loci (Slatkin 1994). Therefore, because all GÉNÉTHON autosomal STR loci have ≥3 alleles, the power to detect significant LD between STR loci should be approximately equivalent.

To assess the distribution of pairwise LD as distinct from multilocus LD, we perform FETs for independence between linked alleles of locus pairs ≤30 cM apart. FETs were implemented by a Monte Carlo procedure, where the hypergeometric probability of the observed table was determined and then compared to hypergeometric probabilities calculated from 17,000 randomly shuffled tables that had the same marginals (Mehta and Patel 1983; Guo and Thompson 1992). The number of times the shuffled table had a hypergeometric probability less than or equal to that of the observed table is taken as the probability that alleles at the loci are independent. The resulting probabilities from these tests for LD are referred to as LDp (linkage disequilibrium probability). The pseudorandom number generator ran1 (Press *et al.* 1992) was used for all permutation-based procedures. Map distance, expressed in centimorgans (cM), was determined from GÉNÉTHON sex-averaged recombination distances (Dib *et al.* 1996).

There are currently no methods available to describe the spatial pattern of LD in an entire genome. Variable marker density and the proportional relationship between the likelihood of LD and interlocus distance present significant challenges to providing an accurate description of the genomic distribution of LD. Such a description must avoid identifying regions with abundant tightly linked markers as exhibiting remarkable concentrations of LD.

In an effort to provide a detailed description of LD within the human genome, a model was developed that corrects for marker density and uses measurements from the data to correct for the relationship between LD and recombination distance. For this model, locus pairs are defined as being "in LD" according to whether their LDp ≤ a cutoff *c*, where *c* is analogous to a multiple test correction. Although this process will misclassify some locus pairs, it simplifies the spatial analysis, and the resulting list of locus pairs provides hypotheses for subsequent empirical evaluation.

Within each 5-cM genomic region a frequency histogram of all pairwise comparisons is produced, based on interlocus distances and with a 0.5-cM bin size (Figure 2a). Within each bin the frequency of locus pairs with a LDp ≤ *c* is determined. The probability of a locus pair having an LDp ≤ *c* for a particular bin was taken as the genome-wide frequency of such pairs, *e.g.*, a 0.1 genome frequency of LDp ≤ *c* for locus pairs within 0.5 cM is taken as the probability. The probability was estimated of observing the same, or more, locus pairs with an LDp ≤ *c* for each 5-cM region of the genome, conditioned on each region's distribution of pairwise distances. Specifi-

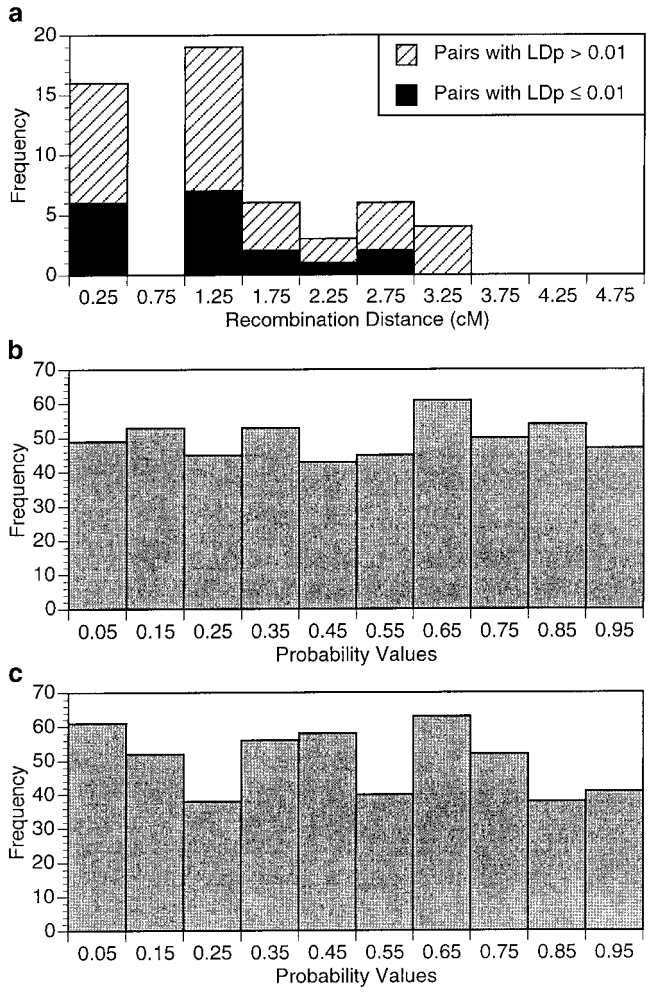

Figure 2.—Testing for clustered LD. (a) Results of spatial distribution assessment of LD in a 5.0-cM window, defined by the STRP locus MogCA, spanning *HLA*. The probability of the observed or more pairs with an LDp ≤ 0.01 for the window is calculated using Equations 1 and 2. Midpoint of 0.5-cM intervals is indicated on *x*-axis. (b) A total of 500 independent windows with all loci having the same heterozygosity and allele frequency distribution. (c) Same as for b, except allele frequency distributions and heterozygosity were randomly drawn from chromosome 1 STR loci.

cally, for each distance bin i within a window w, the binomial probability p_{wi} of the observed or more locus pairs with LDp $\leq c$ is calculated as

$$p_{w_i}(X \geq k) = \sum_{j=k_i}^{n_i} \binom{n_i}{j} p_i^j (1 - p_i)^{n_i - j}, \quad (1)$$

where k_i is the observed number of LDp $\leq c$ locus pairs within distance bin i, n_i is the total number of locus pairs for bin i within the window, and p_i is the probability of a pair having an LDp $\leq c$ for bin i. A novel window statistic ω is computed as

$$\omega_w = \prod_{i, n_i > 0}^{N} p_w, \quad (2)$$

where N (which equals 10 for a 5-cM window) is the number of bins. Small values of ω will correspond to high densities of LD. To establish the probability of the observed ω from window w, it was compared to a distribution of ω calculated from 10^4 random windows (referred to as ω_r values) with an identical distribution of interlocus distances to window w. Random windows were produced by generating the k_i with a pseudorandom number generator (Press *et al.* 1992) using p_i and the n_i from window w over all $n_i > 0$. Values of ω_r were determined from these windows by applying Equations 1 and 2. Observed ω values smaller than all ω_r values were reassessed by increasing the number of random windows to 10^6. The frequency that $\omega_r \leq \omega$ is taken as the probability that the window has the same, or less, abundance of LD than the genome average. To facilitate a graphical inspection of the results, the probabilities from each window in the genomic scan are transformed into χ^2 statistics with 1 d.f. using the standard χ^2 density function and an iterative procedure (Press *et al.* 1992). We subsequently refer to this test as the LD cluster test.

Because the calculation of ω incorporates nonindependent observations, we verified that probabilities from the LD cluster statistic (ω) distribution were approximately uniformly distributed using randomly permuted data. The cluster test was applied to LDp values calculated from 54 randomly shuffled haplotypes under two different scenarios. First, a single locus was selected with no missing data that had, roughly, the median heterozygosity (0.72) and number of alleles (7). The allele frequency distribution at this locus was used for 10^4 loci, evenly spaced 0.25 cM apart, to produce 500 independent 5-cM windows with 20 loci per window. After an initial shuffling of the haplotypes, LDp values were calculated for all pairwise locus comparisons within each window. ω values and their probabilities were estimated for each window using a value of $c = 0.05$. Using the nonparametric Kolmogorov-Smirnov (henceforth KS) test, as implemented in the SAS procedure NPAR1WAY, the distribution is not significantly different from a uniform distribution of the same size ($P = 0.96$; see Figure 2b for a frequency histogram of the probabilities). The second scenario differs from the first only in that the heterozygosity and allele frequency distribution per locus were allowed to vary. Allele frequency distributions were randomly selected from chromosome 1 loci to create a total of 500 independent windows as before. The distribution arising from this second analysis also did not differ significantly from uniform ($P = 0.55$; see Figure 2c for a frequency histogram of the probabilities). Thus, the extent of correlation between comparisons involving the same locus is not significant, and the LD cluster test probability values approximate a uniform distribution.

Population genetics theory predicts that closely linked markers will on average exhibit higher LD (and thus lower LDp values) than loosely linked markers. We test for a relationship between distance (centimorgans) and LDp using the Mantel test for matrix correspondence (Mantel 1967; Sokal and Rohlf 1995) on pairs separated by ≤10 cM. The test compares the two paired matrices of numbers (pairwise recombination distance and pairwise LDp) to assess whether, in this case, small LDp's tend to be associated with small centimorgan distances by multiplying corresponding matrix elements and summing these products across all matrix positions. The observed statistic is then compared to those obtained by randomly shuffling the distance matrix where new positions in the matrix are randomly assigned. The frequency that the shuffled statistic was less than or equal to the observed statistic in 20,000 shufflings is taken as the probability that centimorgan distance and LDp values are independent. To avoid the bias of unresolved map order (0-cM distances), such marker pairs were assigned a distance of 0.1 cM.

Alternative genetic map construction: Concordance between physical and genetic maps has been used to construct highly accurate genomic maps (e.g., see Broman *et al.* 1998). To assess the effect of mapping errors in the recombination linkage map on our results, physical RH map data (Stewart *et al.* 1997) were utilized to obtain alternative estimates of genetic map location. Using version 2 of the Stanford G3 panel (Stewart *et al.* 1997), relative RH map locations for GÉNÉTHON markers within contiguous regions (a group of ≥6 unambiguously linked markers) were determined. RH map locations were plotted against each marker's corresponding GÉNÉTHON map locations. A best-fit line was determined with a parsimonious choice of at most three parameters (X, X^2, and $\log X$) chosen by eye, and then used to predict the alternative, regression-based, genetic map locations for each marker. Linear regression was performed using the GLM procedure of SAS. The sample regressions presented in Figure 3 illustrate the variable relationship between recombination rates and physical distance and show a high degree of concordance in map order. Markers whose map positions were outside the 95% confidence interval of the best-fit line were not considered further. The alternative estimates, obtained for 1438 of 5048 loci, take into account recombination estimates over larger regions and permit estimation of centimorgan distance between markers unresolved on the recombination linkage map.

RESULTS

Evidence for linkage disequilibrium in Europeans: The results of FETs for locus pair independence that were used as an index of LD for 228,955 locus pairs are presented in Figure 4 as a function of recombination distance (centimorgans). In Figure 4, a and b, we present a frequency histogram of pairs with a test outcome of LDp ≤ 0.05. In accordance with population genetics theory, the percentage of pairs with LDp (P value for departure from allele independence) values in this range is highest in the shortest intervals, 0–0.5, 0.5–1.0, and 1.0–1.5 cM. Moreover, in the interval 0–0.5 cM, the majority of these LDp ≤ 0.05 pairs exhibit LDp values ≤ 0.01 (Figure 4b).

The pattern of LDp *vs.* centimorgan distance between STRP loci within short (≤3.5-cM) intervals prescribes a linear relationship between the percentage of pairs with small LDp values and the inverse of centimorgan distance (1/cM) between test loci (Figure 4c). For 0.5-cM intervals from 0 to 3.5 cM, the relationship is highly significant ($r^2 > 0.99$; $P < 10^{-6}$ for LDp ≤ 0.01; Figure 4c), suggesting a strong proportionality of centimorgans and LDp for loci 3.5 cM apart. This result is not depen-

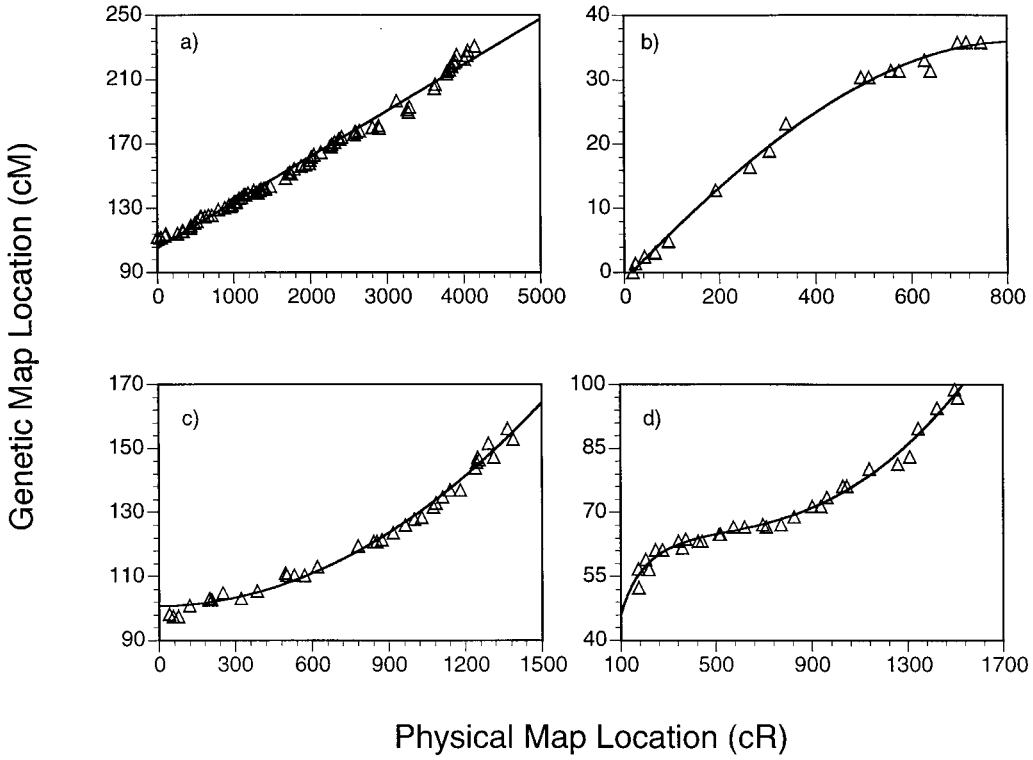

Figure 3.—An alternative map using concordance between physical and genetic maps. Presented are four representative sample regressions from portions of chromosomes 3 (a), 3 (b), 11 (c), and 18 (d).

dent upon the influential point at 4 cM, because its removal has little impact on the regression relationship ($r^2 > 0.97$; $P < 10^{-3}$). To assess the potential confounding effect of map errors in the recombination linkage map on this relationship, we analyzed alternative values predicted from concordance between GÉNÉTHON and Stanford RH maps as described in materials and methods and illustrated in Figure 3. The alternative estimates are based on the physically mapped markers and provide order and non-0-cM estimates between markers unresolved on the GÉNÉTHON map. The reanalysis (Figure 4c) affirms the proportionality of inverse centimorgans to the likelihood of LD.

A plot of mean LDp for locus pairs separated by discrete (1-cM) recombination distances is presented (Figure 4d). A relationship between the centimorgan distance separating a locus pair and their corresponding LDp value was tested for using the Mantel test (Mantel 1967; Sokal and Rohlf 1995). Performing this analysis on loci separated by ≤10 cM (82,846 locus pairs) revealed a significant correlation between LDp value and centimorgan distance ($P < 0.001$; Mantel 1967; Sokal and Rohlf 1995). To assess the limit of this correlation and thus the limit of LD in Europeans, the GÉNÉTHON dataset was titrated to successively exclude locus pairs in the 0–1.0, 0–2.0, 0–3.0, etc., ranges until the Mantel test yielded probabilities that were >0.05. This analysis provides the conservative estimate that, for samples with $N = 54$, the upper threshold for correlation between centimorgans and LDp is 4.0 cM.

Because low *P* values for multiple statistical tests can result from chance alone (as opposed to LD), a multiple test correction is necessary. However, due to the large number of FETs performed, a Bonferroni-based estimate of *c* (the LDp cutoff) is much less than the resolution of the Monte Carlo FET. Accordingly, an alternative method was employed to obtain an estimate of *c* below which the majority of locus pairs are likely to be in LD. Specifically, the linear relationship of LDp *vs.* 1/cM (Figure 4c) was used to identify the range of LDp values for which the majority of pairs were in authentic LD. By titrating probabilities in 0.01 *P*-value intervals, it was determined that while the percentage of locus pairs with LDp ≤ 0.01 are highly correlated with 1/cM, those locus pairs in higher *P*-value intervals (*e.g.*, 0.01 < LDp ≤ 0.02) are not (results not shown) suggesting that only locus pairs with an LDp ≤ 0.01 are predominantly in authentic LD. Interestingly, the LDp ≤ 0.01 relationship would predict that the empirical limit of LD approximates 5.5 cM from the GÉNÉTHON-based data set. This number, obtained by solving the regression equation (Figure 4c) for the distance at which the proportion of pairs in LD is equal to the expectation of 1%, is remarkably consistent with the distance in Figure 4d, where the mean LDp *vs.* centimorgan curve asymptotes at ~6.5 cM with the background expectation of 0.5. These results offer strong statistical support for implicating LD for the majority of locus pairs separated by 4 cM whose LDp ≤ 0.01. Thus, for subsequent analyses, the cutoff $c = 0.01$ was used. Out of 36,382 locus pairs within 4 cM of each other, 1452 (4%) have an LDp ≤ 0.01. A list of these locus pairs and their LDp values is available at either jcsmr.anu.edu.au/~glenys/humgen/data.htm or rex.nci.nih.gov/RESEARCH/basis/lgd/front_page.htm/.

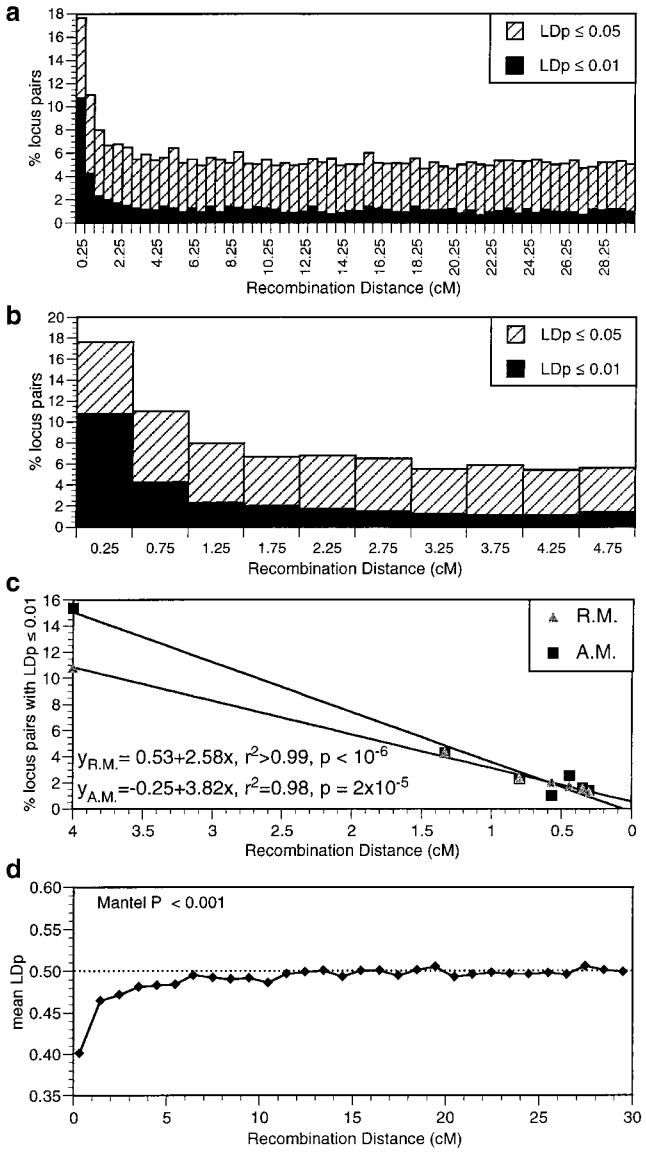

Figure 4.—Relationship between LD and recombination distance. (a) Histogram showing frequency (percentage) of locus pairs with small LDp values. The midpoint of each 0.5 cM is bin listed on the *x*-axis. (b) Same as for a, but restricted to loci within 5 cM of each other. (c) Plot, equation, and statistics for the percentage of locus pairs with LDp ≤ 0.01 *vs.* 1/cM, where cM values are the interval midpoints from a. Data are from the recombination linkage map (R.M.) and the alternative map (A.M.). (d) Centimorgan distance between STRP markers and average LDp values between markers within each centimorgan interval. Mantel *P*, probability for the relationship between LDp and distance from the Mantel test for pairs within 10 cM.

Consistent with the suggested dependence of probabilities from exact tests on heterozygosity (Slatkin 1994), a significant, but very small, correlation between LDp and heterozygosity was detected. If heterozygosity was influential in the power of loci to detect significant LD, then for loci within 0.5 cM of each other, a difference in heterozygosity should be apparent between those loci with an LDp ≤ 0.01 and the remaining loci. Using a KS test, we reject the null hypothesis that the two groups of loci have the same heterozygosity distributions ($P < 10^{-3}$). The median heterozygosity value for loci with an LDp ≤ 0.01 (∼0.73; 1381 loci) is greater than the median value for the remaining loci (∼0.72; 2986 loci). To assess the proportion of variation in LDp values accounted for by heterozygosity, a multiple regression was performed on ∼500 independent locus pairs within 0.5 cM of each other, but separated from all other locus pairs by at least 5 cM. Taking heterozygosity at loci A and B as independent variables and LDp as the dependent variable, the analysis indicated that heterozygosity at the two loci accounts for <0.04% of the variance in LDp.

Linkage disequilibrium is heterogeneously distributed throughout the genome in Europeans: Different evolutionary forces may produce different spatial patterns of LD in the genome. The null hypothesis of spatial homogeneity of LD was initially tested by comparing the LDp distribution of individual chromosomes to the rest of the genome. For example, the LDp distribution (from locus pairs within 4 cM of each other) of chromosome 1 was compared to the LDp distribution from the rest of the genome (produced by pooling the LDp values from chromosomes 2–22). Because differences in LDp values could arise from differences in marker density, the datasets representing a chromosome and the genome were matched for the distribution of interlocus distances. To illustrate this, if 5% of all comparisons within 4 cM on chromosome 1 were between loci separated by 0.3 cM, while for the genome set this value was 8%, locus pairs were randomly sampled without replacement from the genome set to achieve a proportion of 5%. Performing the nonparametric KS test on the 22 comparisons indicates seven chromosomes (2, 5, 6, 12, 13, 15, and 18) had probabilities ≤0.05, with chromosomes 2, 5, and 18 having more LD than the genome average and the other chromosomes having less LD. Of these, chromosomes 2 ($P = 0.0007$), 15 ($P = 0.0001$), and 18 ($P = 0.0013$) are significant after correcting for multiple tests using the Bonferroni procedure (Sokal and Rohlf 1995). We therefore reject the null hypothesis that LDp values, and thus LD, are uniformly distributed in the genome. A comparison of the chromosome heterozygosity distributions for the loci in each of the above sets, again using the KS test, failed to detect any chromosomes significantly different from the genome average. These results suggest that the nonuniform distribution of LD at the chromosomal scale does not result from variation in locus power (as represented by heterozygosity) to detect LD.

Given apparent heterogeneity of LD in the genome, and prior to analyzing the entire genome, we evaluated the effectiveness of the LD cluster test on the human leukocyte antigen (HLA). The HLA region on chromosome 6 includes several loci previously reported to dis-

play LD as a consequence of selective pressure on epistatic loci in the region (Bodmer 1986; Klein 1986; Hedrick 1994; Trowsdale 1995; Foissac *et al.* 1997) and associated LD among linked STR loci (Carrington *et al.* 1998). HLA loci are highly polymorphic (like STRP loci) and play an important role in foreign antigen processing and presentation to T-cell receptor molecules on immune lymphocytes. Because of insufficient GÉNÉTHON marker density, the same CEPH families used in this study were genotyped with an additional six STRP markers. The subsequent analysis of this region proved significant even after adjusting for multiple tests from seven windows ($P = 0.002 <$ adjusted significance $P = 0.007$). We interpret this positive result to affirm the approach in detecting clusters of LD throughout the genome.

The results from the cluster detection analysis using the GÉNÉTHON dataset are shown in Figure 5. To appraise the contribution of variation in marker density to these results, 500 independent (nonoverlapping) windows were analyzed. The results indicate that marker density accounts for <2% of the variance in window probabilities. Two approaches were employed to judge the significance of the results in Figure 5. First, a standard Bonferroni multiple test correction was determined using the total number of windows over the entire genome ($N = 4575$). Using this correction, region 7 on chromosome 16 is significant ($P = 3 \times 10^{-6} <$ corrected 5% significance level $P = 1.1 \times 10^{-5}$). However, this approach is overly conservative in part due to the Bonferroni correction itself (Rice 1989; Rothman 1990) and partly because of correlations between overlapping windows. Second, a less stringent (but still conservative) approach was employed by using the HLA region as a benchmark. This region has previously been shown to exhibit extensive LD and it may therefore serve as a lower bound for identifying other such significant regions (Bodmer 1986; Klein 1986; Hedrick 1994; Trowsdale 1995; Foissac *et al.* 1997; Carrington *et al.* 1998). This approach yields nine additional regions distributed across seven chromosomes (Table 1) and the centimorgan length of regions between STRPs in LD ranges from 2.2 to 6.2 cM (excluding HLA). Genes identified in these regions, which may be influenced by or have a role in developing the initial LD, are listed in Table 1.

Clustering of LD may also stem from the accumulation of loci with high power to detect LD, genotyping errors, or mapping errors. Although clusters may arise from rare concentrations of loci with high power to detect LD, no difference was detected between levels of heterozygosity at loci with an LDp ≤ 0.01 within the clusters and the general distribution of heterozygosity. Such a relationship was tested for in two ways. First, a KS test was used to compare the distribution of heterozygosity from all loci in the clusters with an LDp ≤ 0.01 to the distribution of heterozygosity for all other loci ($P = 0.17$). Second, a two-tailed sign test was conducted to evaluate whether the frequency of loci with either heterozygosities less than, or greater than, the median value from all loci ($\sim$0.72) was unusually high over all clusters or for each cluster separately. None of the sign tests were statistically significant. Only region 8 exhibited a small probability ($P = 0.03$) for detecting five loci with heterozygosities greater than the median. These results suggest that the clusters do not stem from rare accumulations of informative markers.

Mapping errors could contribute to the detection of clusters either by genotyping errors, underestimating recombination fractions resulting from the number of meioses sampled, or incorrect ordering of loci. Genotyping errors can exert a significant impact on interlocus distance estimates, causing an overestimation of total map length (Brzustowicz *et al.* 1993). This tends to decrease the significance of a region in the LD cluster test. High concordance between the genetic and physical maps (>99.5%; Hudson *et al.* 1995) suggests that the genotyping error rates in the GÉNÉTHON dataset are probably very small. However, if genotyping errors increase with increasing STR fragment size, then concentrations of loci with large mean fragment size might lead to clustered LD. A comparison of mean fragment sizes of loci in LD within the clusters to the rest of the genome failed to support this hypothesis ($P = 0.27$ from the KS test).

To assess the effect of underestimating recombination rates, distance estimates were divided by a factor of 2 or 4 with the same division of window size. Using these modified data, a LD cluster test was performed, using the genome-wide averages from the unaltered analysis, for each window within the regions presented in Table 1. In an attempt to test whether the regions would still exhibit a density of LD comparable to HLA from our original analysis, all windows within the regions were compared to the *P* value 0.002. Region 6 had a single window fulfilling these criteria in the fourfold compressed analysis ($P = 0.002$) and a comparable result in the twofold analysis ($P = 0.007$). Region 9 had probabilities <0.05 in both the two- and fourfold analyses, while region 5 had $P < 0.05$ for the twofold analysis only. The HLA and other regions (1, 2, 3, 4, 7, 8, and 10) were not significant in both analyses ($P > 0.5$). While these results show that the LD cluster test is sensitive to mapping errors, one region (6) is somewhat robust in the face of such errors.

The alternative map was also used to evaluate the contribution of individual errors in recombination distance to the results. The relationship between the physical and genetic maps was used to identify markers discordant between the two maps. These markers were eliminated, and the physical map location of the remaining markers was used to reestimate genetic map locations. Consequently, the alternative map incorporates regional estimates of recombination fraction. Phys-

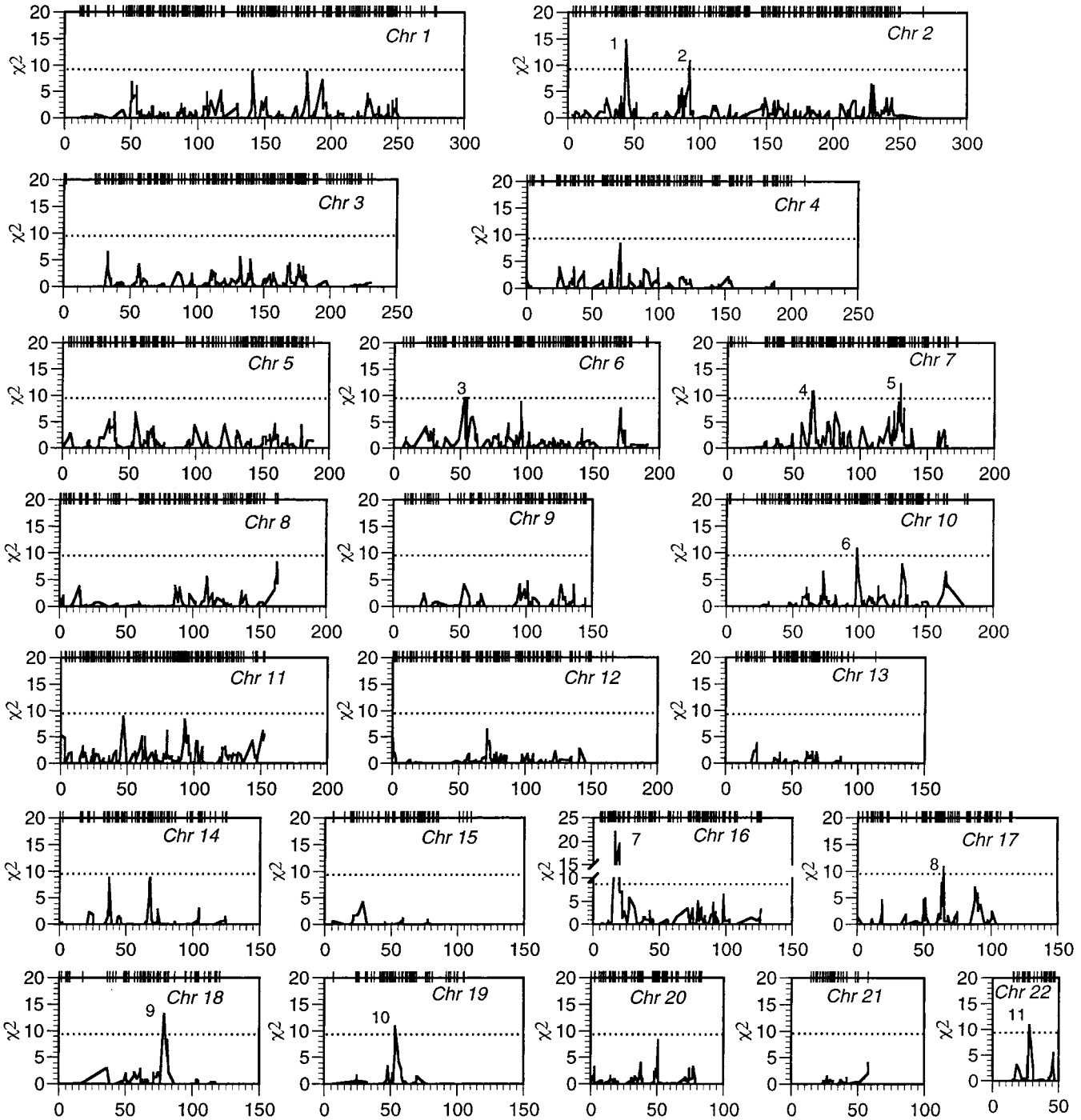

Figure 5.—Distribution of linkage disequilibrium (LD) throughout the human genome in a 5-cM sliding window.The *x*-axis is the chromosomal location of the anchor locus (the first "upstream" locus in the window) based on the recombination linkage map; the *y*-axis shows the χ^2 with 1 d.f. for the probability of the observed or greater LD within each 5.0-cM window. Tick marks above plots indicate window positions. Information on the numbered regions is summarized in Table 1. The horizontal dotted line at the height of the HLA cluster is used as a reference to identify remarkable regions.

ical map estimates of intermarker distances and marker order may have higher error rates than the genetic map (Deloukas *et al.* 1998), potentially jumbling the correct order. Insufficient alternative map data prevent us from assessing all but regions 2 and 5. The results provide significant support for regions 2 and 5 after adjusting

TABLE 1

Summary of significant linkage disequilibrium regions

Region	Chromosome loc	cM loc	STRP limit	Max LD interval (cM)	No. loci	No. pairs in LD/ no. loci in LD	*P*	Genes in region
1	2p25	43.7	D2S2373–D2S2170	4.3	8	8/8	0.0001	APOB, HADHB, hFKBP-12, KHK
2[a]	2p15	92.1	D2S303–D2S145	2.8	11	10/8	0.001	ACTG2, ANX4, GFPT, MAD, P47, SM protein G, TGFA
3	6p21-p22	48.4	MIB–D6S1666	10.4	12	12/12	0.002	HLA
4	7p13	60	D7A2497–D7S2427	4.8	20	9/14	0.001	AEBP1, AMPH, BLVR, GCK, GLI3, INHBA, OGDH, POU-domain factor-1, PRKCSH, RALB, RAMP3, UDP
5[a]	7q31-q32	128.4	D7S650–D7S530	6.2	23	18/18	0.0005	ARF5, PTPRZ
6	16p13	15.6	D16S3020–D16S3069	6.2	15	14/12	3×10^{-6}	BCMA, GSPT1, hNR2A, KAI1, MHC2TA, P5CDh, PRM1, XMP
7	17q21-q22	63	D17S1787–D17D943	3.3	24	14/15	0.001	ARF4L, BRCA1, CDC27, CNP, COL1A1, DLG2, DLG3, DLX4, ETV4, GFAP, GIP, GRN, HOXB1, HOXB13, HOXB5, IGFBP4, ITGB2B, ITGB3, KRT10, MOX1, MYL4, NDP52, NGFR, RPL27, SLC4A1, SP2, TOP2A, UBTF
8	18q21	78.8	D18S1152–D18S68	4.2	8	4/5	0.0003	FECH, LRP1
9	19q13	55.9	D19S225–D19S213	2.2	5	4/4	0.001	ATP4A, CAPN4, CD22, CEBPG, COX6B, GPI, HPN, PEPD
10	22q12	27.7	D22S424-D22S1173	4.9	5	3/5	0.001	CSF2RB, HMOX1, IL2RB, LGALS1, MYH9, RAC2, TST

Regions with probabilities less than or equal to that of HLA (Figure 5). Chromosome loc, cytogenetic location; cM loc, genetic map location of the first locus. We define the borders of significant regions as the limit of consecutive windows whose *P* value was ≤ 0.05, and with at least one window with a *P* value less than or equal to that of HLA. Loci with an LDp ≤ 0.01 closest to borders of these regions are shown. Max LD interval, the largest distance between a locus pair in LD; *P*, the probability from the LD cluster test. Gene symbols for genes mapping to the region were ascertained from the National Center for Biotechnology Information (NCBI) gene map using the indicated STRP loci (www.ncbi.nlm.nih.gov/genemap/).

[a] Significant regions using the alternative map centimorgan estimates after adjusting for multiple tests.

for multiple comparisons ($P = 0.0032$ for both regions), even though the data for these regions were still incomplete. As such, the nine regions that were detected with the limited haplotype sample ($N = 54$) indicate LD signals of potential import in the ancestry of this European population.

DISCUSSION

The exploration of the distributional properties of LD in Europeans was conducted at three levels: level 1, a genome-wide average description of the relationship between locus pairs in LD and the recombination distance separating them (Figure 4, a–d); level 2, a chromosome scale analysis to determine whether LD is uniformly distributed across the genome; and level 3, a detailed regional analysis for locus clusters that depart from the genomic background LD (as in level 1; Figure 5). The level 1 analysis indicated considerable sporadic LD among loci linked by ≤4.0 cM, the proportion of which was inversely related to centimorgan distance (Figure 4c). The level 2 analysis showed that LD is heterogeneous in its genomic distribution in a marker density-independent fashion. The level 3 5.0-cM sliding-window analysis revealed nine genomic regions with clustered LD greater than or equal to that observed for HLA, which include the known genes listed in Table 1.

We detected a striking proportionality between LD and inverse recombination fraction (Figure 4). This relationship indicates that while LD occurs between loci within 5 cM of each other, the majority of these pairs cluster within the shortest distance interval. However, the linearity between the proportion of locus pairs with small LDp values and 1/cM has limitations. Over extremely short distances the relative contribution of mutation to the decay of LD is larger, reducing the role for recombination and thus impacting on the 1/cM result, while at extended distances the regression relationship will predict negative estimates for the proportion of loci in LD, which is biologically implausible.

There are two broad potential evolutionary origins for the observed LD: genetic drift or natural selection (Ohta and Kimura 1969; Hartl and Clark 1990). Because European populations have had relatively large effective population sizes ($N_e \geq 10{,}000$), are known to have expanded rapidly in recent centuries during agricultural development, and have not experienced appreciable recent founder effects (Takahata *et al.* 1992; Takahata 1993; Ayala 1995; Ayala and Escalante 1996; von Haeseler *et al.* 1996), mutation and genetic drift appear unlikely explanations (Slatkin 1994). Moreover, a plausible consequence of drift-derived LD is a similar LDp distribution on all chromosomes. Yet, both our chromosome scale and sliding-window analyses indicate that the spatial pattern of LD in the genome is significantly nonuniform.

Our ascertainment, in the level 3 analysis, of 10 genomic regions exhibiting remarkable concentrations of loci in LD is plausibly an underestimate because the two multiple test corrections used are conservative. The most stringent of these, the Bonferroni correction, identifies only region 6 as being significant. However, we reject the strict Bonferroni correction as a general guideline for interpreting the results of this analysis because it tends to produce type II error, particularly when multiple tests are not independent of each other, as is the case in this analysis (Rice 1989; Rothman 1990). Furthermore, it may be argued that our alternative strategy of using the HLA region to define a lower benchmark is also overly restrictive because this region is known for its high level of LD (Bodmer 1986; Klein 1986; Hedrick 1994; Trowsdale 1995; Foissac *et al.* 1997).

Although several classes of errors might have contributed to the spatial pattern of LD, our analyses did not support their involvement. The impact of errors in regional distance estimates and locus order for the chromosome scale analysis should be small. In contrast, the impact of regional underestimation of recombination on the sliding-window analyses is potentially severe. Despite this, region 6 on chromosome 16 still exhibited low probabilities. Furthermore, the effect of ordering errors was assessed using the alternative map for two regions (2 and 5). That both these regions were significant supports the authenticity of the remaining regions as representing clustered LD in the European genome.

A further potential confounding factor is variation in the power of loci to detect LD. Although heterozygosity was significantly higher for loci with an LDp ≤0.01 relative to the remaining loci, it accounts for <0.04% of the variance in LDp values. Further, we did not detect differences in heterozygosity concordant with the LDp distributions of chromosomes or between loci in the regions defined in Table 1 and the remainder of the genome. The potentially confounding influence of variable informativeness may be substantially reduced in these data by the consistently high heterozygosity prevalent among the GÉNÉTHON STR loci.

The nonuniform pattern of LD in the genome is consistent with the operation of natural selection. However, selective explanations for linkage disequilibrium have been proposed previously only for the HLA region (Bodmer 1986; Klein 1986; Hedrick 1994; Trowsdale 1995). Thus, the absence of an explicit test to discriminate between neutral and selective origins of LD at a genomic scale prevents a formal conclusion regarding the evolutionary origin of the detected LD.

The LD genome screen described here offers a new perspective on the organization of endemic genetic variation in the human genome. Although the haplotype sample size is limited ($N = 54$), and thus a sizable portion of LD is potentially undetected in Europeans, the analysis of 5048 loci was nonetheless informative in describing a background level for LD in Europeans as

well as identifying specific genomic locales where LD appears elevated. The background LD might also be useful in LD association studies that are increasingly being applied to locate genes contributing to heritable disease and phenotypes.

We thank Cecile Fizames from CEPH for assistance in obtaining the genotype data, several anonymous reviewers, Andy Clark, Simon Easteal, George Nelson, Clay Stephens, and Sue Wilson for helpful comments on this article. We thank the Frederick Biomedical Supercomputing Center for their assistance. The content of this article does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government. This project was funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract no. NO1-CO-56000.

LITERATURE CITED

Ayala, F. J., 1995 The myth of Eve: molecular biology and human origins. Science **270:** 1930–1936.

Ayala, F. J., and A. A. Escalante, 1996 The evolution of human populations: a molecular perspective. Mol. Phylogenet. Evol. **5:** 188–201.

Begovich, A. B., G. R. McClure, V. C. Suraj, R. C. Helmuth, N. Fildes *et al.*, 1992 Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. J. Immunol. **148:** 249–258.

Bennett, B. M., and P. Hsu, 1960 On the power function of the exact test for the 2×2 contingency table. Biometrika **47:** 393–398.

Bodmer, W. F., 1986 Human genetics: the molecular challenge. Cold Spring Harbor Symp. Quant. Biol. **51:** 1–13.

Briscoe, D., J. C. Stephens and S. J. O'Brien, 1994 Linkage disequilibrium in admixed populations: applications in gene mapping. J. Hered. **85:** 59–63.

Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White and J. L. Weber, 1998 Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am. J. Hum. Genet. **63:** 861–869.

Brzustowicz, L. M., C. Merette, X. Xie, L. Townsend, T. C. Gilliam *et al.*, 1993 Molecular and statistical approaches to the detection and correction of errors in genotype databases. Am. J. Hum. Genet. **53:** 1137–1145.

Carrington, M., D. Marti, J. Wade, W. Klitz, L. Barcellos *et al.*, 1999 Microsatellite markers in complex disease: mapping disease-associated regions within the human Major Histocompatibility Complex, pp. in *Microsatellites: Evolution and Applications*, edited by D. B. Goldstein and C. Schlötterer. Oxford University Press, Oxford.

Chakraborty, R., and K. M. Weiss, 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. USA **85:** 9119–9123.

Deloukas, P., G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund *et al.*, 1998 A physical map of 30,000 human genes. Science **282:** 744–746.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

Ewens, W. J., and R. S. Spielman, 1995 The transmission / disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet. **57:** 455–464.

Foissac, A., B. Crouau-Roy, S. Fauré, M. Thomsen and A. Cambon-Thomsen, 1997 Microsatellites in the HLA region: an overview. Tissue Antigens **49:** 197–214.

Guo, S. W., and E. A. Thompson, 1992 Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics **48:** 361–372.

Gyapay, G., J. Morissette, A. Vignal, C. Dib, C. Fizames *et al.*, 1994 The 1993–94 Généthon human genetic linkage map. Nat. Genet. **7:** 246–339.

Hartl, D. L., and A. G. Clark, 1990 *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA.

Hastbacka, J., A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. **2:** 204–211.

Hedrick, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.

Hedrick, P., 1994 Evolutionary genetics of the major histocompatibility complex. Am. Nat. **143:** 945–964.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

Hudson, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.

Hudson, T. J., L. D. Stein, S. S. Gerety, J. Ma, A. B. Castle *et al.*, 1995 An STS-based map of the human genome. Science **270:** 1945–1954.

Jorde, L. B., 1995 Linkage disequilibrium as a gene-mapping tool. Am. J. Hum. Genet. **56:** 11–14.

Kaplan, N. L., P. O. Lewis and B. S. Weir, 1994 Age of the delta F508 cystic fibrosis mutation. Nat. Genet. **8:** 216–218.

Kaplan, N. L., W. G. Hill and B. S. Weir, 1995 Likelihood methods for locating disease genes in nonequilibrium populations. Am. J. Hum. Genet. **56:** 18–32.

Klein, J. (Editor), 1986 *Natural History of the Major Histocompatibility Complex*. John Wiley and Sons, New York.

Klitz, W., J. C. Stephens, M. Grote and M. Carrington, 1995 Discordant patterns of linkage disequilibrium of the peptide-transporter loci within the HLA class II region. Am. J. Hum. Genet. **57:** 1436–1444.

Laan, M., and S. Paabo, 1997 Demographic history and linkage disequilibrium in human populations. Nat. Genet. **17:** 435–438.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

Lewontin, R. C., and K. Kojima, 1960 The evolutionary dynamics of complex polymorphisms. Evolution **14:** 458–472.

Mantel, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer Res. **27:** 209–220.

Martin, M. P., A. Harding, R. Chadwick, M. Kronick, M. Cullen *et al.*, 1998 Characterization of 12 microsatellite loci of the human MHC in a panel of reference cell lines. Immunogenetics **47:** 503.

McLellan, T., L. B. Jorde and M. H. Skolnick, 1984 Genetic distances between the Utah Mormons and related populations. Am. J. Hum. Genet. **36:** 836–857.

Mehta, C. R., and N. R. Patel, 1983 A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Am. Stat. Assoc. **78:** 427–434.

Moonsamy, P. V., W. Klitz, M. G. Tilanus and A. B. Begovich, 1997 Genetic variability and linkage disequilibrium within the DP region in the CEPH families. Hum. Immunol. **58:** 112–121.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics **63:** 229–238.

Peterson, A. C., A. Di Rienzo, A. E. Lehesjoki, A. de la Chapelle, M. Slatkin *et al.*, 1995 The distribution of linkage disequilibrium over anonymous genome regions. Hum. Mol. Genet. **4:** 887–894.

Press, W. H., S. A. Teukovsky, W. T. Vettering and B. P. Flannery, 1992 *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, United Kingdom.

Raymond, M., and F. Rousset, 1995 An exact test for population differentiation. Evolution **49:** 1280–1283.

Rice, W. R., 1989 Analyzing tables of statistical tests. Evolution **43:** 223–225.

Rothman, K. J., 1990 No adjustments are needed for multiple comparisons. Epidemiology **1:** 43–46.

Slatkin, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137:** 331–336.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*. W. H. Freeman and Company, New York.

Stephens, J. C., D. Briscoe and S. J. O'Brien, 1994 Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am. J. Hum. Genet. **55:** 809–824.

Stephens, J. C., D. E. Reich, D. B. Goldstein, H. D. Shin, M. W. Smith *et al.*, 1998 Dating the origin of the CCR5-Δ32 AIDS resistance allele by the coalescence of haplotypes. Am. J. Hum. Genet. **62:** 1507–1515.

Stewart, E. A., K. B. McKusick, A. Aggarwal, E. Bajorek, S. Brady *et al.*, 1997 An STS-based radiation hybrid map of the human genome. Genome Res. **7:** 422–433.

Takahata, N., 1993 Allelic genealogy and human evolution. Mol. Biol. Evol. **10:** 2–22.

Takahata, N., Y. Satta and J. Klein, 1992 Polymorphism and balancing selection at the major histocompatibility complex loci. Genetics **130:** 925–938.

Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd *et al.*, 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271:** 1380–1387.

Trowsdale, J., 1995 "Both man and bird and beast": comparative organization of MHC genes. Immunogenetics **41:** 1–17.

von Haeseler, A., A. Sajantila and S. Paabo, 1996 The genetical archaeology of the human genome. Nat. Genet. **14:** 135–140.

Weissenbach, J., G. Gyapay, C. Dib, A. Vignal, J. Morissette *et al.*, 1992 A second-generation linkage map of the human genome. Nature **359:** 794–801.

Wiehe, T., and M. Slatkin, 1998 Epistatic selection in a multilocus Levene model and implications for linkage disequilibrium. Theor. Popul. Biol. **53:** 75–84.

Communicating editor: A. G. Clark