Am. J. Hum. Genet. 67:926–935, 2000

Consistent Long-Range Linkage Disequilibrium Generated by Admixture in a Bantu-Semitic Hybrid Population

James F. Wilson[1,2] and David B. Goldstein[1]

[1]Department of Biology, University College London, London; and [2]Department of Zoology, University of Oxford, Oxford

Both the optimal marker density for genome scans in case-control association studies and the appropriate study design for the testing of candidate genes depend on the genomic pattern of linkage disequilibrium (LD). In this study, we provide the first conclusive demonstration that the diverse demographic histories of human populations have produced dramatic differences in genomewide patterns of LD. Using a panel of 66 markers spanning the X chromosome, we show that, in the Lemba, a Bantu-Semitic hybrid population, markers ⩽~21 cM apart have a significantly greater tendency to show LD than do unlinked markers. In three populations with less evidence of admixture, however, excess LD disappears >2 cM. Moreover, analysis of Bantu and Ashkenazi populations as putative parental populations of the Lemba shows a significant relationship between allele-frequency differentials and the LD observed in the Lemba, which demonstrates that much of the excess LD is due to admixture. Our results suggest that demographic history has such a profound effect on LD that it will not be possible to predict patterns a priori but that it will be necessary to empirically evaluate the patterns in all populations of interest.

Introduction

Case-control association studies are increasingly the method of choice in efforts to map genes underlying complex traits (Risch and Merikangas 1996). It remains unclear, however, as to how densely markers must be distributed throughout the genome in order to allow reliable detection of association with causal variants (Collins et al. 1999; Kruglyak 1999). In particular, in the consideration of variants relevant to common disease, the genetic distance over which significant linkage disequilibrium (LD) occurs will determine the appropriate spacing of markers, whereas the consistency of LD at a given genetic distance will determine the number of markers of a given type that are required within each interval.

Uncertainty about optimal marker spacing is due, in part, to our ignorance of the nature of the genetic variation that influences common disease. In particular, recent mutations will tend to have more LD than will older ones, and it is not clear whether the common diseases are influenced more by common or by rare alleles (Kruglyak 1999; Wright et al. 1999). Uncertainty is greatly compounded, however, by our ignorance of the distribution of LD in human populations, for any class of variants. It should be noted that, even when exhaustive single-nucleotide polymorphism (SNP) maps are developed, the interpretation of association studies will still require detailed knowledge of LD. In this case, the pattern of LD will determine the genetic distance over which false signals of causation may be generated by the association of a candidate SNP with a linked causal variant. This complication would apply equally to tests, using case-control designs, of the role of specific variants in candidate genes. For these reasons, interest has focused on the measurement and prediction of levels of LD in human populations. These patterns of association, however, are influenced both by demographic factors, which affect the entire genome, and by genetic factors, such as mutation rates and selection, which influence particular genomic regions (Freimer et al. 1997; Wright et al. 1999). Uncertainties concerning the demographic histories of human populations, however, make it difficult to accurately predict patterns of LD, even in the case of neutrality. In a recent theoretical study, for example, Kruglyak (1999) noted that, under simplified demographic assumptions, useful LD would not, in most human populations, extend beyond ~3 kb. The little evidence that is available, however, indicates much more complicated demographic scenarios for most populations, involving both rapid expansions and bottlenecks (Reich and Goldstein 1998; Collins et al. 1999).

Optimal design and interpretation of association studies, therefore, require empirical study of LD between pairs of markers widely distributed in multiple genomic regions. A description of such "background"

Received June 21, 2000; accepted for publication August 8, 2000; electronically published August 28, 2000.

Address for correspondence and reprints: Dr. David B. Goldstein, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London, NW1 2HE, United Kingdom. E-mail: d.goldstein@ucl.ac.uk

0002-9297/2000/6704-0015$02.00

LD would allow observed associations to be evaluated in the context of the distribution of LD in the relevant population. At present, however, most studies of background LD are limited in either the number of genomic regions considered (Laan and Pääbo 1997), the number and type of populations studied (Huttley et al. 1999), or the range of genetic distances represented (Goddard et al. 2000, but see Taillon-Miller et al. 2000). It is noteworthy that, even in the model populations targeted in many association studies (e.g., Ashkenazi Jews and Finns), the assumptions of relative homogeneity and limited disequilibrium $\gtrsim$0.5 cM have not been confirmed through systematic genetic evaluation. In fact, in the case of Finland, preliminary data suggest both genetic heterogeneity (Kittles et al. 1998) and the presence of some long-range LD (Peterson et al. 1995). What is critically needed is a consistent experimental framework, for measurement of LD across large genomic regions, that can be applied to multiple populations with different demographic histories.

In this article, we concentrate on admixture, arguably the most important demographic factor that is ignored in theoretical treatments seeking to predict patterns of LD in human populations. It is well known that population substructuring (or stratification) can generate spurious signals in association studies, and various methods have been proposed to identify it (Pritchard and Rosenberg 1999) and take it into account in statistical analyses (Reich and Goldstein, in press). Even if no stratification were present, however, historical admixture between differentiated populations could result in significantly elevated LD over large genomic regions, for many generations (Chakraborty and Weiss 1988; Stephens et al. 1994; McKeigue 1998). In fact, it has even been suggested that, in some populations, the LD generated by admixture would allow genome scans with only one marker every 10 cM or so, more than three orders of magnitude less than the theoretical suggestion of Kruglyak (1999). These predictions about patterns of LD in admixed populations, however, have yet to be validated empirically.

In this study, we report on the analysis of 66 microsatellite markers specifically designed to allow formal assessment of the effect that admixture has on LD. The markers were selected to provide dense and, as far as possible, uniform coverage of a broad range of genetic distances. Since this panel was developed to assess the effect of recent admixture, we have concentrated attention on relatively large genetic distances, and the coverage within the interval 0–1 cM is therefore very uneven. Collectively, these markers provide 2,145 pairwise observations of LD in multiple regions throughout the X chromosome (fig. 1).

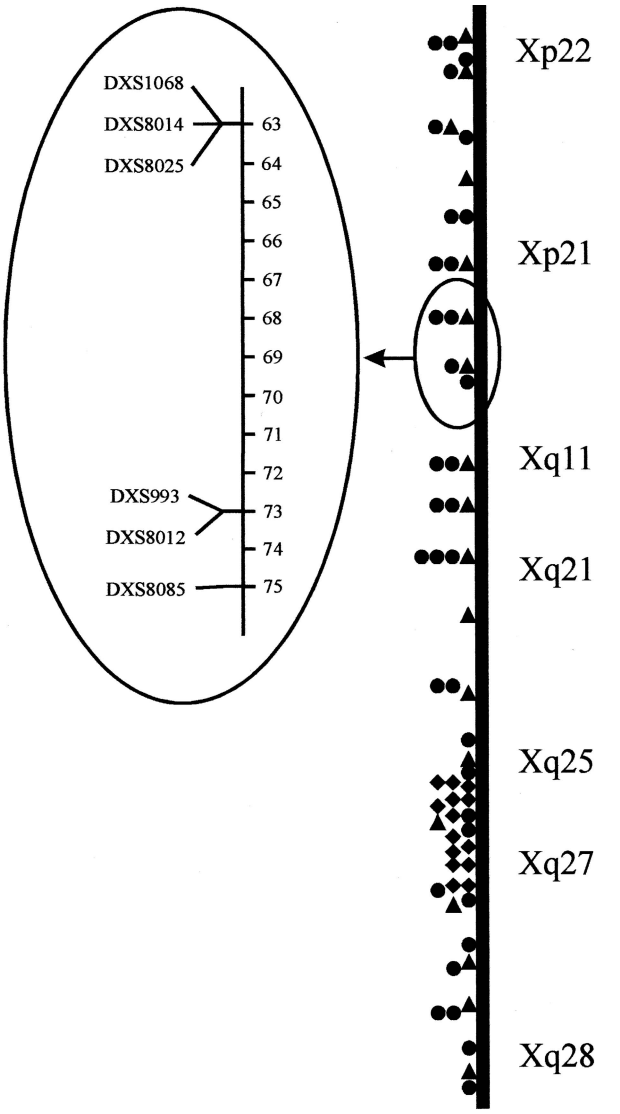

Figure 1 Distribution of the 67 markers in the admixture-LD panel across the X chromosome. Note that DXS1047 is not reported in this study. Approximate cytogenetic locations are indicated. Genetic distances (in cM) are according to the Généthon map. The X-chromosome ABI Prism linkage mapping–panel markers (▲) are spaced every ~10 cM across the chromosome; flanking these are usually two microsatellites within 1 cM (●) (see detail). Together, these provide multiple pairwise comparisons over genetic distances of 0–1 cM and multiples of 10 cM. Distances between 1 and 10 cM and their multiples are similarly covered by a series of microsatellites (♦) at Xq25–27.

Material and Methods

Samples

All DNA samples were collected from paternally unrelated males. Lemba subjects originated from Sekhukuneland, in Mpumalanga, South Africa; Bantu-speakers were from various chieftainships in the Pretoria area; and mixed-caste Ashkenazi Jews were from Tel Aviv

(Thomas et al. 2000). Ethiopians were sampled in Addis Ababa and consisted mainly of Amharic and Oromo speakers from the Wollo and Shewa provinces.

Genotyping

The admixture LD marker panel was chosen as described in figure 1. Markers in addition to the X-chromosome ABI Prism linkage-mapping panel (PE Biosystems) were chosen from The Genome Database, and some primers were redesigned from GenBank sequences, as were primers for a novel marker 7.5 kb from DXS1203 in PAC 455H14. Multiplex PCR was performed on 65 dinucleotide loci (from the Généthon linkage map [Dib et al. 1996]), one tetranucleotide ($GATA_n$) locus (from the Marshfield map [Broman et al. 1998]), and the novel (CA_n) microsatellite, in 13 kits, details of which are available from the authors.

In brief, a 10 × primer master mix was made in advance; the PCR master mix was then made up of primer master mix (1.5–5.0 pmol for each primer), 10 × Super *Taq* PCR buffer 1 (HT Biotechnologies), 0.2 mM of each dNTP (Advanced Biotechnologies), Super *Taq* (HT Biotechnologies) and *Taq*Start (Clontech) (premixed in a 2:1 ratio, neat *Taq*: neat *Taq*Start), and, finally, dH_2O to a total of 10 μl per reaction. Meanwhile, ~10–20 ng of DNA were pipetted into a 96-well microtiter plate. After being vortexed, the master mix was aliquoted into microtiter plate–strip lids. After a centrifugation step, the samples were cycled in a Perkin-Elmer 9700 PCR machine (PE Biosystems). Thermal profiles were 38 cycles of 30 s at each of 95°C, 55°C, and 72°C, with a 4-min 95°C initial denaturation step and a 10-min 72°C final extension. In some cases, a touchdown procedure was used, decreasing the annealing temperature by 0.5°C/cycle for the first eight cycles. The Prism panel markers were cycled according to the manufacturer's instructions. PCR products were diluted and pooled, prior to being loaded onto a 96-lane ABI 377 sequencer (PE Biosystems). In this way, 3–18 markers were loaded within each lane, allowing 288–1,728 genotypes per gel to be called. The average number of loci per lane was 10. Size calling was performed using GENESCAN software. A sample from a control individual was run at least twice on each gel to standardize for gel-to-gel shifts in migration. The locus names are DXS7103, 1232, 8027, 8087, 8036, 984, 1067, 8009, 8061, 8085, 1204, 8092, 8099, 996, 8078, 1212, 1036, 1205, 8081, 1223, 1053, 8013, 8098, 1220, 8086, 8014, 8038, 1206, 8073, 1062, 1192, 1211, 1203, 8068, 1193, 8025, 8032, 8057, 8072, 8094, 6801, 8067, 8012, 994, 8059, 8105, 8028, 8082, 1227, 990, 986, 987, 993, 1073, 8091, 1106, 1047, 1001, 1068, 1214, 8055, 8051, 8043, 1060, 1226, 991, and the novel locus in PAC 455H14. Note that DXS1047 was not typed in the Ashkenazim and so was not used in the analysis.

Analysis

LD was measured by a test of independence between alleles at pairs of loci (Slatkin 1994), which is only sensitive to marginal frequencies. This extension of Fisher's exact test for R × C contingency tables was implemented using the Arlequin program (Schneider et al. 1997). The probability of finding a table with the same marginal totals that has a probability equal to or less than that of the observed table was obtained using a Markov chain to efficiently explore the space of all possible randomly shuffled tables (Guo and Thompson 1992). These P_{LD} values were calculated by first dememorizing the chain for 10,000 steps and running the chain in batches, until the error of the P value was <.001. Contingency tables testing for interactions between either genetic distance or the product of interpopulation allele frequency difference (δ) values at two loci (i.e., $\delta_1\delta_2$) and P_{LD} were evaluated using χ^2. Corrections for multiple comparisons involving the windows covering different genetic distances were made using the Dunn-Sidak sequential correction (Sokal and Rohlf 1995), $P' = 1 - (1 - \alpha)^{1/k}$, where k is the rank of the P value, from highest to lowest, and where $\alpha = .05$. The correction is conservative in our case, since the windows are overlapping and, therefore, nonindependent. Composite δ values for each locus were calculated as the sum of all allele × allele δ values of like sign. The log-linear test for a three-way interaction between P_{LD}, distance (in cM), and $\delta_1\delta_2$ was implemented in STATISTICA.

Results

Populations

The admixture-LD panel was genotyped in the Lemba ($n = 96$), a Bantu-Semitic hybrid population, and in three other populations, all of which have less evidence of admixture. The Lemba are a southern African group who speak a variety of Bantu languages and claim Jewish ancestry. Y-chromosome analysis (Spurdle and Jenkins 1996; Thomas et al. 2000) indicates that ~68% of the Lemba Y chromosomes are Semitic and that 32% are Bantu in origin. Although it was not possible to distinguish between an Arabic contribution and a Jewish one, the high frequency of the Cohen modal haplotype, a putative Hebrew signature, may indicate a specifically Judaic component in the Lemba gene pool. mtDNA studies (Soodyall et al. 1996) found no evidence of female-mediated admixture. Both as putative parental populations of the Lemba and as examples of populations with less evidence of admixture, samples of non-

Lemba South African Bantu speakers ($n = 86$) and Ashkenazi Jews ($n = 80$) were chosen. The latter are also of interest because of their frequent use in genetic epidemiology (Wright et al. 1999) and because they may have undergone some admixture with Europeans. As an alternative second parental population, a sample of Ethiopians ($n = 77$) (described in the Material and Methods section) were chosen. The Y chromosomes in this sample were more similar to the non-Bantu Lemba chromosomes than were those of the Ashkenazim (data not shown). Only the first 34 markers listed in the Material and Methods section were genotyped in the Ethiopian sample.

Extent and Consistency of LD

LD was measured using an extension of Fisher's exact test, to give P_{LD}. In total, the proportion of marker pairs in significant LD ($P_{LD} < .05$) in the Lemba (13.8%) is twice that in Ashkenazi Jews (7.0%), Bantu (7.7%), and Ethiopians (6.4%). To investigate the tendency of marker pairs to be in LD as a function of genetic distance, we compared, using contingency tables, the LD in a given distance interval versus that between unlinked marker pairs. The tables compare the counts of significant and nonsignificant P_{LD} values, in 29 windows—each 6 cM wide and centered at 1-cM intervals along the chromosome (0–5 cM, 1–6 cM, etc.)—versus the numbers of significant and nonsignificant unlinked pairs (distance $\geqslant$50 cM) (fig. 2). The comparison to unlinked pairs was intended to control for any inherent tendencies of the loci to be nonrandomly associated. We developed this approach rather than using the more common Mantel test because we are not interested in demonstrating a trend in the relationship of LD with distance. Rather, we want to identify the point at which marker pairs show LD indistinguishable from that seen for unlinked pairs.

The Lemba show elevated LD at very large genetic distances, compared with unlinked markers: there is a significant excess of pairs in LD, out to the 19–24-cM interval (fig. 2). This contrasts sharply with the Bantu and Ashkenazim, in whom excess LD extends only to the 1–6-cM window, and with the Ethiopians, who show only weak LD in the 0–5-cM window. Using $P_{LD} < .01$ as a significance threshold does not alter the results. After correction for multiple comparisons, P values for 17 of the 20 intervals out to 19–24 cM are significant in the Lemba.

To better localize the point at which excess LD disappears in the Bantu and Ashkenazi Jews, we constructed contingency tables with 1-cM windows; only the 0-, 1-, and 2-cM tables are significant in each population (not shown). In the Ethiopians, none of the intervals contain a significant excess of LD. In the Lemba, excess LD extends to the 19–24-cM window, but, because of the lower number of pairs in the 19–24-cM interval, it is not possible to use 1-cM windows. However, with 2-cM windows, no tables beyond 21–22 cM are significant. LD is thus maintained across genetic distances that are an order of magnitude greater in the Bantu-Semitic hybrid population than they are in the other populations. It should be noted, however, that the comparison with unlinked pairs demonstrates that the LD is not independent of genetic distance, as would be observed in a substructured population sample.

In addition to the genetic distance over which LD occurs in the Lemba, the consistency of the LD is also remarkable. In the Bantu and Ashkenazi Jews, in the 0–2-cM interval, 46% and 38%, respectively, of pairs are in LD; but, in the Lemba, one and a half times as many pairs—63%—show significant LD in this interval (fig. 3). The proportion of pairs in LD rapidly decreases with increasing genetic distance, in all populations except the Lemba: for markers at 3–23 cM, 20% of pairs show significant LD in the Lemba, versus only 4%–7% in the other three populations. Thus, the Lemba have increased LD at most genetic distances, but LD is increased only in the 0–2-cM interval in the Bantu and Ashkenazi Jews; and, in these latter two populations, 38% and 25%, respectively, of marker pairs separated by exactly 2 cM show LD, versus 0% in the Ethiopians. Fewer marker pairs were genotyped in the Ethiopians; however, only 18% of pairs in the 0–2-cM interval show LD. Very little allelic association is observed at any genetic distance in the Ethiopians.

Even for unlinked markers, the Lemba show an increased level of LD: almost twice as many pairs (10%) are in disequilibrium, compared with what is seen in the other three populations (6%), which could reflect differences due to either recent admixture, drift, or substructure. Since the extent of LD was compared with that between unlinked markers, it is therefore all the more remarkable that an excess extends out to ~21 cM. Furthermore, since the admixture is thought to be principally male mediated, X-linked markers will show reduced LD, compared with autosomal markers, although this is offset, since the X chromosome has both less opportunity to recombine and a lower effective population size.

Lemba LD: Generated by Admixture?

To the extent that the Lemba LD has resulted from admixture between Semitic and Bantu peoples, the δ value between the parental populations should be predictive of the LD observed (Stephens et al. 1994). In particular, in the case of two alleles at each locus, at a given genetic distance, $\delta_1\delta_2$ should be linearly related to the LD (Chakraborty and Weiss 1988; Briscoe et al.

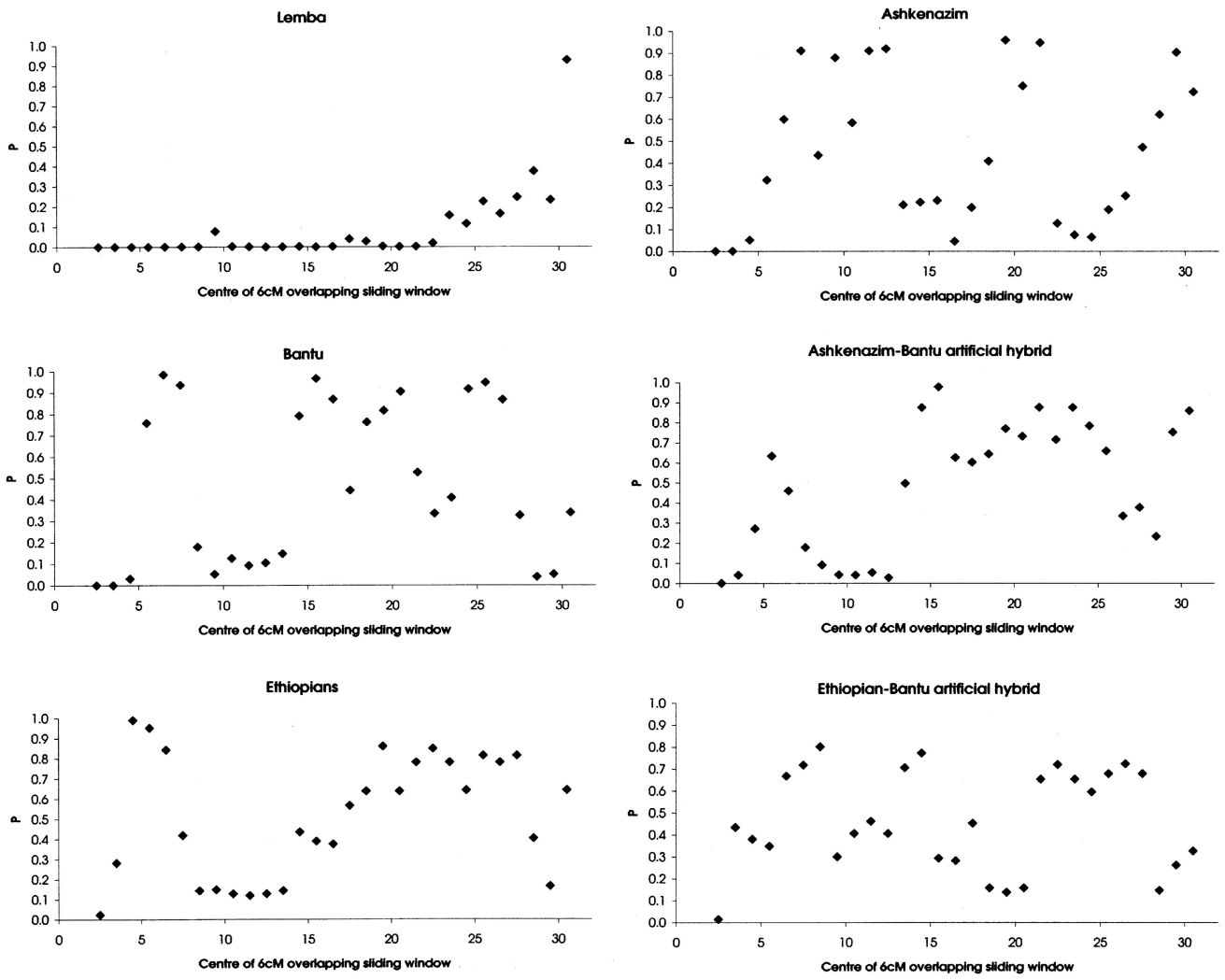

Figure 2 Extent of LD. The excess of significant LD is tested for by a comparison of pairs of markers within a given range of genetic distances versus unlinked pairs of markers. For each point in the graphs, a contingency table is evaluated in which the first column contains the numbers of significant and nonsignificant marker pairs with genetic distances in a window delimited by $r - 2.5$ and $r + 2.5$, where r is plotted on the X-axis. In all cases, the second columns contain the numbers of significant and nonsignificant P_{LD} values for all unlinked marker pairs.

1994). We used the composite δ for multiallelic systems that has been suggested by Shriver et al. (1997) (see the Material and Methods section). The predicted relationship between δ and LD provides a framework both to investigate whether LD is admixture generated and for the identification of populations that have given rise to hybrid groups. We used two combinations of putative parental populations for the Lemba: Ashkenazi-Bantu and Ethiopian-Bantu.

In relating δ values to LD, we exclude marker pairs in the 0–1-cM interval, since these show considerable LD in the parental populations, some of which will have been transmitted to the Lemba, obscuring the signal of admixture-generated LD. In fact, 66% of significant Lemba marker pairs at 0–1 cM are also significant in the Bantu or Ashkenazi Jews, compared with only 21% of those at 2–16 cM. When the Ashkenazi-Bantu δ values are used, the proportion of significant Lemba marker pairs at 2–16 cM increases in higher $\delta_1\delta_2$ classes (fig. 4). Of the pairs with $\delta_1\delta_2 > .3$, 38% are in LD, but, of those with $\delta_1\delta_2 < .1$, only 15% are in LD. This effect is not observed at distances >17 cM, where the proportion of significant pairs in different $\delta_1\delta_2$ classes is 10%–15%. To assess the significance of this difference, we built a contingency table by counting the number of significant ($P_{LD} < .05$) and nonsignificant marker pairs with $\delta_1\delta_2$ val-

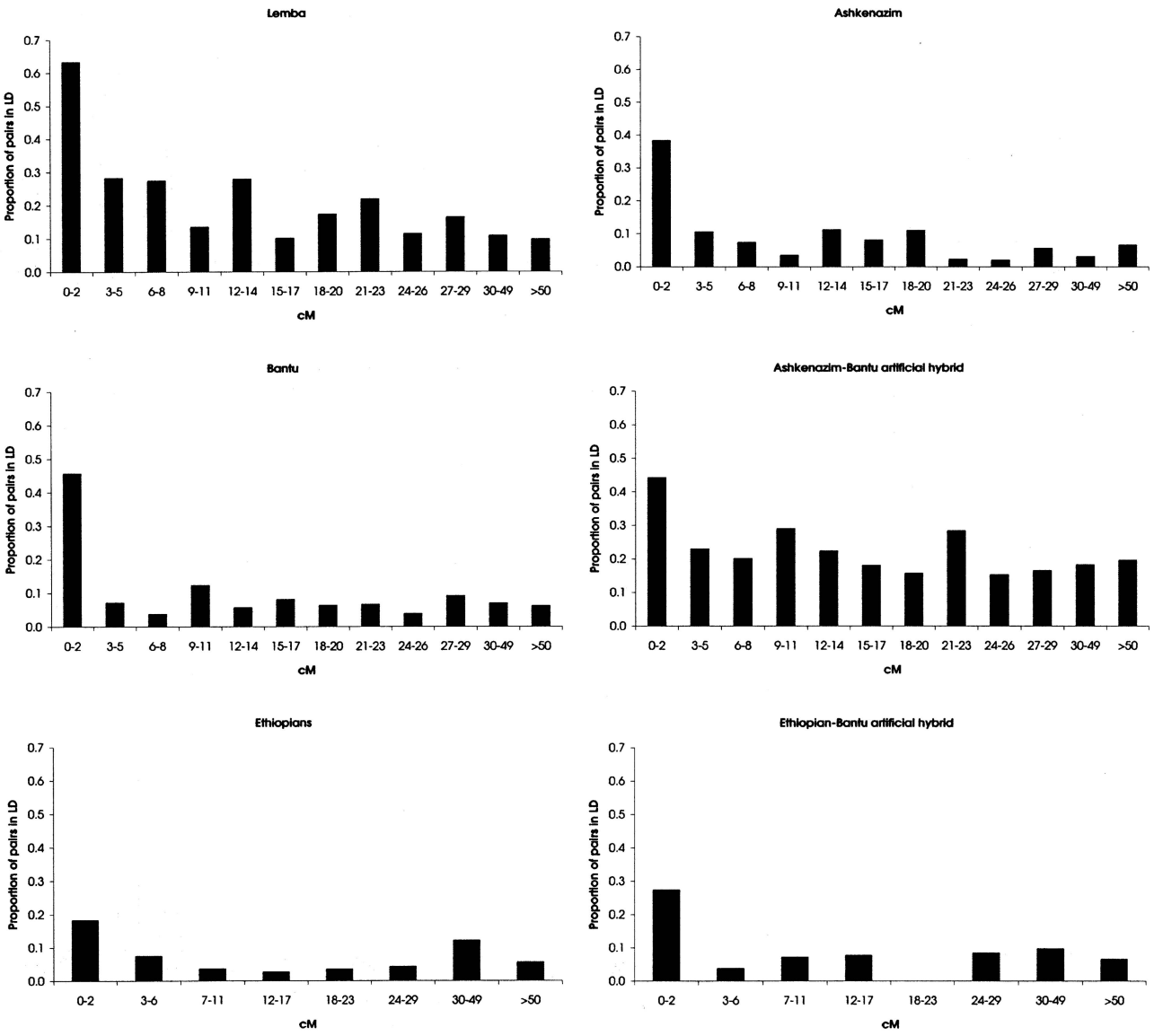

Figure 3 Consistency of LD at different genetic distances: histograms of the proportion of locus pairs, in various distance classes, with $P_{LD} < .05$. Note that the distance classes are wider in the Ethiopians, since fewer markers were genotyped.

ues above and below the mean (.17). At 2–16 cM, significantly more pairs in the high-$\delta_1\delta_2$ category show LD ($P = .04$). For unlinked pairs, this contingency table is not significant ($P = .2$): there is little effect of $\delta_1\delta_2$ on the probability of showing significant LD, even though both the $\delta_1\delta_2$ values and LD are influenced by allele frequencies. When the 2–16-cM contingency table is evaluated using the Ethiopians as the second parental population, there is no significant interaction between $\delta_1\delta_2$ and P_{LD}, either with the Ashkenazi-Bantu threshold of .17 or with the Bantu-Ethiopian mean $\delta_1\delta_2$ value. Using multiway tables, we also tested whether the degree of association between P_{LD} and Ashkenazi-Bantu $\delta_1\delta_2$ is significantly different for linked and unlinked marker pairs; however, the three-way interaction is not significant ($P = .3$).

Artificial Hybrid Populations

To better evaluate the predicted effect that Bantu-Semitic admixture has on LD, we constructed an artificial hybrid population by randomly sampling 45 X chromosomes from each of two putative parental populations of the Lemba: the Ashkenazi Jews and the Bantu. In the resulting artificial hybrid population, 20.5% of all marker pairs are in LD—one and a half times as many

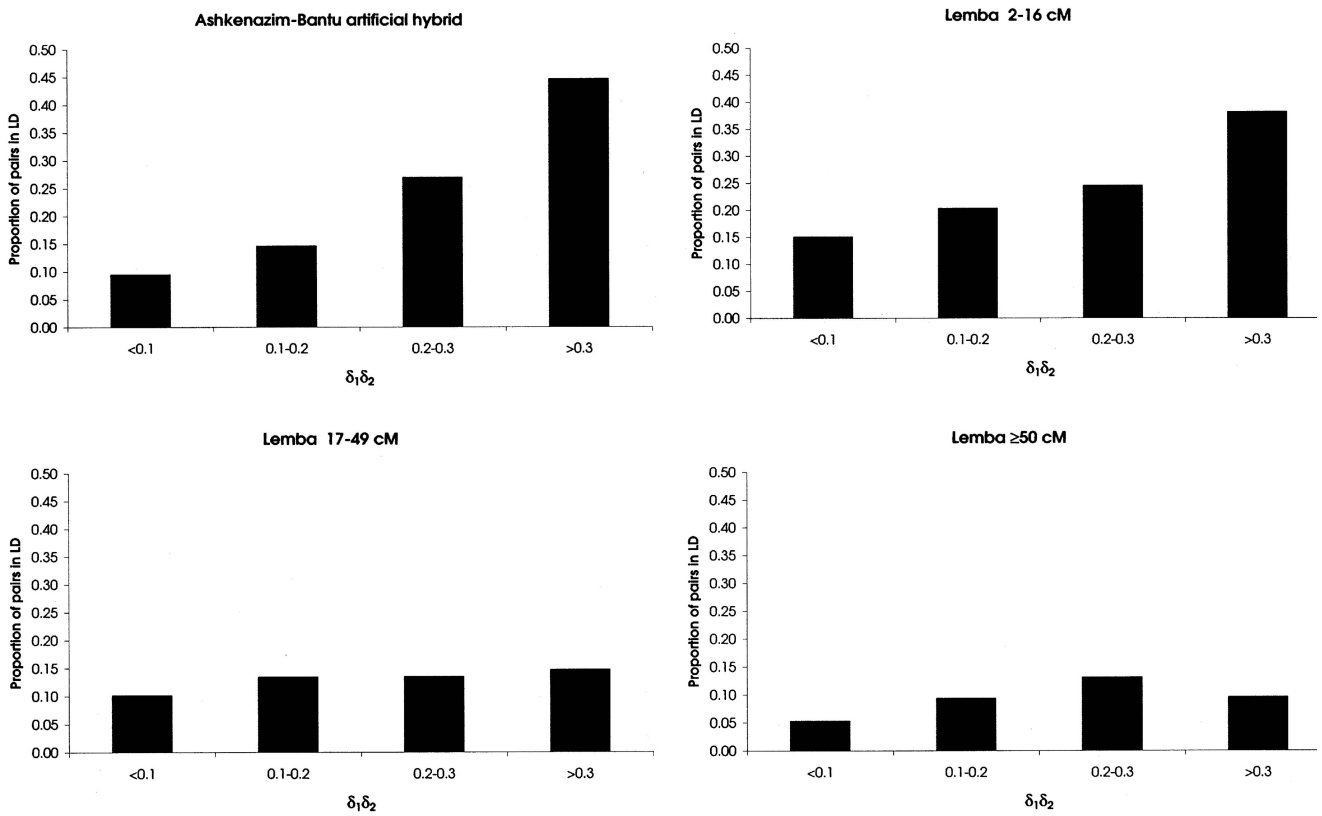

Figure 4 Proportion of pairs in LD in various Ashkenazi-Bantu $\delta_1\delta_2$ classes. Lemba pairs are divided into three genetic-distance classes; however, all pairs are shown for the Ashkenazi-Bantu artificial hybrid population.

as in the Lemba. As expected, the substructuring results in disequilibrium largely independent of genetic distance, with 19.6% of unlinked markers in LD (fig. 3). Also as expected, a higher proportion of pairs are in LD in higher $\delta_1\delta_2$ classes, regardless of the genetic distance: 45% of pairs with $\delta_1\delta_2 > .3$ are in LD, versus only 9% of pairs with $\delta_1\delta_2 < .1$ (fig. 4). Significantly more pairs with above-average $\delta_1\delta_2$ values show LD ($P < 10^{-18}$). Since it takes only three generations to reduce by half the LD at 20 cM (and since it takes a single generation to do so for unlinked markers), comparison of the LD levels in both the artificial hybrid and the Lemba populations suggests either that at least some admixture must have occurred extremely recently in the latter or that there is some substructure. The significant difference between partially linked and unlinked loci, however, rules out substructure as the sole source of the LD in the Lemba. In contrast, when an artificial Bantu-Ethiopian hybrid is created in the same manner, very little LD is generated (fig 3). Only 7.5% of pairs are in LD, a frequency that is comparable to that in the parental populations, which indicates that Ethiopian-Bantu differentiation is not sufficient to produce the disequilibria observed in the Lemba. Moreover, there is little relationship between $\delta_1\delta_2$ and LD in this artificial hybrid (data not shown). Combining (*a*) the fact that Bantu-Ethiopian δ values do not predict Lemba LD and (*b*) the general lack of differentiation between these putative parentals, we conclude that the Ethiopians are not a good representation of the non-Bantu parental group of the Lemba.

Ancestral LD Remains Detectable in a Stratified Population

Because substructure creates LD at all genetic distances, no overall tendency to disequilibrium above that between unlinked pairs is observed in the artificial hybrid populations (fig 2). Interestingly, however, excess LD can still be detected at the smallest genetic distances. This excess is significant only in the Ashkenazi-Bantu artificial population, in which 44% of pairs are in LD in the 0–2-cM interval, compared with 20% for unlinked pairs (figs. 2 and 3). The LD in the 0–2-cM interval must have two sources: LD inherited from the parental populations (which is the LD useful in association studies) and that created de novo by admixture/sample stratification (which is the spurious LD in mapping studies). However,

when multiple pairs of markers in the 0–2-cM interval are used, this ancestral LD remains detectable over and above the stratification LD observed with Bantu-Semitic levels of subpopulation differentiation. Furthermore, the proportion of pairs showing extreme LD ($P_{LD} < 10^{-4}$) is 10-fold higher for pairs at 0–1 cM (17%), compared with unlinked pairs (1.7%). Case-control association studies may thus overcome spurious signals of association, by means of designs using multiple marker pairs to assess the magnitude of LD generated by stratification and by correcting association statistics accordingly (Pritchard et al. 2000; Reich and Goldstein, in press).

Variance and LD

To investigate the relationship between the variability of a locus and LD, we evaluated contingency tables comparing significant and nonsignificant pairs with repeat-count variance above and below the mean (6.2). In the Lemba, there is, between 0 and 16 cM, a significant positive interaction between the mean variance and P_{LD} ($P = .005$). This may be because high-variance loci have greater power to show significant LD, or it could reflect the positive relationship between the mean repeat-count variance in the parental populations and the δ values ($P < 10^{-10}$). In this same interval, no such relationship between variance and LD is seen in the Bantu, Ethiopians, and Ashkenazi Jews. In fact, in the Jews and Ethiopians, for tightly linked markers (0–3 cM), the opposite trend is observed: low-variance locus pairs have a tendency to show more-significant LD; in this case, the LD is probably older, and mutation probably has contributed to the reduction of LD at the high-variance loci.

Discussion

Earlier work has claimed to have identified admixture-generated LD, most recently at the Duffy locus in African Americans (Parra et al. 1998; Lautenberger et al. 2000). The observation of a significant excess of haplotypes carrying together the alleles commonly found in European populations suggests that admixture has indeed contributed to the observed LD. However, another recent study of the Duffy locus in the parental African population has revealed strong evidence of selection (Hamblin and Di Rienzo 2000), which has probably generated LD around the locus; an unknown amount of the LD observed in African Americans may be such ancestral LD inherited from the African parental population. To provide conclusive evidence of the effect that demography has on LD, it is critical to analyze multiple genomic regions (Freimer et al. 1997). Our analysis demonstrates the presence of a significant excess of LD between markers $\leq$21 cM apart, compared with that between unlinked markers in the admixed Lemba. Moreover, analysis of δ between the Bantu and Ashkenazi Jews shows a significant predictive effect on the LD observed in the Lemba, demonstrating that the LD was generated by Bantu-Semitic admixture. The elevated LD at unlinked markers suggests that some of the Lemba LD may be due to very recent admixture, but, since there is no relationship with Ashkenazi-Bantu δ values, this admixture possibly involved a different parental population. In any case, the difference between linked and unlinked intervals (compared with that in our artificial hybrid populations) rules out structure as the only source of excess LD.

In the Ashkenazi Jews, significant excess LD, compared with what is seen for unlinked markers, is observed out to 2 cM, as may be expected in light of both admixture with European populations and possible founder effects (although improbably tight bottlenecks may be required to generate this level of LD [Kruglyak 1999]). In the southern Bantu, the presence of excess LD out to 2 cM contrasts with data indicating that, at the *CD4, PAH,* and *DM* loci, African populations show less LD than do non-African populations (Tishkoff et al. 1996, 1998; Kidd et al. 2000). The observed LD may be due to a population-specific event, such as an older admixture event, since there is, in southern Africa, both genetic and linguistic evidence of admixture between the indigenous Khoisan peoples and the incoming Bantu (Cavalli-Sforza et al. 1994). Moreover, the admixture likely occurred during the past 20 generations, which would be recent enough for considerable LD to persist at <2 cM today. In contrast with these populations, Ethiopians have a much lower amount of LD, emphasizing the powerful influence that unknown aspects of a population's demographic history has on the pattern of LD.

Under a set of simplifying assumptions concerning demographic history, it has been predicted that, for a case-control study of typical sample size, useful LD would be unlikely to extend to >3 kb (Kruglyak 1999). Contrary to this expectation, we have observed significant LD in 25%–38% of marker pairs separated by 2 cM (~2,000 kb on average) in the Bantu and Ashkenazi Jews, two populations without evidence of extensive admixture. It should be noted that the sizes of our samples ($n = 77$–96) correspond to only moderately sized case-control studies, unless the number of individuals required for detection of association between a causal variant and the trait that it influences is large (see Kruglyak 1999). The discrepancy between expected and observed patterns of LD suggests that very few, if any, real populations will match the assumptions—and, therefore, the predictions—of such a simplified demographic history. In the case of the Lemba, a population with evidence of a more extreme demographic history, we

observe LD across genetic distances more than three orders of magnitude greater than that predicted from simple demographic assumptions. Despite the considerable LD generated by admixture, however, analysis of an artificial hybrid population shows that the ancestral LD between tightly linked markers is detectable over and above the spurious associations created by stratification. The profound effect that demographic history has on the background LD that we have documented in this study indicates that, in general, it will not be possible to predict patterns of LD a priori but, rather, that it will be necessary to empirically evaluate the patterns in all populations of interest. The observed patterns of LD do, however, present an important (Tishkoff et al. 1996, 1998) and largely untapped source of information regarding human evolutionary history.

Acknowledgments

We thank the Centre for Genetic Anthropology, University College London, for providing samples; Fiona Gratrix and Raphaelle Chaix, for technical assistance; and David Reich and Adrian Hill, for useful discussions. This work was supported, in part, by a Royal Society grant to D.B.G.

Electronic-Database Information

The URLs for data in this article are as follows:

Généthon, http://www.genethon.fr (for dinucleotide loci)
Genome Database, The, http://gdbwww.gdb.org/ (for markers)

References

Briscoe D, Stephens JC, O'Brien SJ (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. J Hered 85:59–63

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63:861–869

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Chakraborty R, Weiss KM (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

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 96:15173–15177

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154

Freimer NB, Service SK, Slatkin M (1997) Expanding on population studies. Nat Genet 17:371–373

Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216–234

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669–1679

Huttley GA, Smith MW, Carrington M, O'Brien SJ (1999) A scan for linkage disequilibrium across the human genome. Genetics 152:1711–1722

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. Am J Hum Genet 66: 1882–1899

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. Am J Hum Genet 62:1171–1179

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

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

Lautenberger JA, Stephens JC, O'Brien SJ, Smith MW (2000) Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. Am J Hum Genet 66:969–978

McKeigue PM (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 63:241–251

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63: 1839–1851

Peterson AC, Di Rienzo A, Lehesjoki AE, de la Chapelle A, Slatkin M, Freimer NB (1995) The distribution of linkage disequilibrium over anonymous genome regions. Hum Mol Genet 4:887–894

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

Reich DE, Goldstein DB (1998) Genetic evidence for a Paleolithic human population expansion in Africa. Proc Natl Acad Sci USA 95:8119–123

Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol (in press)

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) Arlequin version 1.1: a software for population genetic analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60:957–964

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

Sokal RR, Rohlf FJ (1995) Biometry. WH Freeman, New York

Soodyall H, Vigilant L, Hill AV, Stoneking M, Jenkins T (1996) mtDNA control-region sequence variation suggests multiple independent origins of an Asian-specific 9-bp deletion in sub-Saharan Africans. Am J Hum Genet 58:595–608

Spurdle AB, Jenkins T (1996) The origins of the Lemba "black Jews" of southern Africa: evidence from p12F2 and other Y-chromosome markers. Am J Hum Genet 59:1126–1133

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

Taillon-Miller P, Bauer-Sardina I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok PY (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat Genet 25: 324–328

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, Goldstein DB (2000) Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba—the "black Jews of southern Africa." Am J Hum Genet 66:674–686

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271: 1380–1387

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402

Wright AF, Carothers AD, Pirastu M (1999) Population choice in mapping genes for complex diseases. Nat Genet 23: 397–404