

Heather E. Collins-Schramm · Bill Chima
Darwin J. Operario · Lindsey A. Criswell
Michael F. Seldin

Markers informative for ancestry demonstrate consistent megabase-length linkage disequilibrium in the African American population

Received: 20 January 2003 / Accepted: 18 March 2003 / Published online: 3 June 2003

© Springer-Verlag 2003

Abstract Admixture mapping is a potentially powerful tool for mapping complex genetic diseases. For application of this method, admixed individuals must have genomes composed of large segments derived intact from each founding population. Such segments are thought to be present in African Americans (AA) and should be demonstrable by examination of linkage disequilibrium (LD). Previous studies using a variety of polymorphic markers have variably reported long-range LD or rapid decay of LD. To further define the extent and characteristics of LD caused by admixture in the AA population, the current study utilized a set of 52 diallelic markers that were selected for large standard variances between putative representatives of the founder populations. LD was examined in over 250 marker-pairs, including linked markers from four different chromosomal regions and an equal number of matched unlinked comparisons. In the representative founder populations, strong LD was not observed for markers separated by more than 10 kb. In contrast, results indicated significant LD ($P < 0.001$, $D' > 0.3$) in AA over large genomic segments exceeding 10 centiMorgans (cM) and 15 megabases (Mb). Only marginally significant LD was present between unlinked markers in this population, suggesting that choosing appropriate levels of significance for admixture mapping can minimize false positive results. The ability to detect LD for extended chromosomal segments in AA decayed not only as a function of the distance between markers, but also as a function of the standard variance of the markers. This examination of several

genomic segments provides strong evidence that appropriate selection of informative markers is a crucial prerequisite for the application of admixture mapping to the AA population.

Introduction

The usefulness of applying linkage disequilibrium (LD) to identify or confirm the importance of genetic loci in the inheritance of simple Mendelian and complex genetic disease has received much attention (Tsui et al. 1986; Risch and Merikangas 1996; Terwilliger and Weiss 1998). LD that extends for long distances is ideal for these genome-wide mapping endeavors, as fewer markers will then be required to ensure that at least one is in LD with a genetic locus of interest. LD can be created by mutation, selection or demographic factors, including population bottlenecks and admixture (Terwilliger and Weiss 1998). However, because LD decays rapidly as a function of recombination, the mutation events that contribute to disease are unlikely to generate long-range LD, except for the case of new rare Mendelian disorders. Similarly, long-range LD would be created by only very strong and recent selection events, which are unlikely to have occurred for many prevalent diseases. Populations that have undergone recent bottlenecks have received attention for their potential for having long intervals of LD, although the strength of this LD for segments greater than 1 centiMorgan (cM) is not clear (Lonjou et al. 1999; Jorde et al. 2000; Mohlke et al. 2001). Another, and perhaps the best potential source of long range LD may be recent admixture between divergent populations that differ substantially in the frequency of specific marker and disease susceptibility alleles. It has been suggested that recent admixture between ethnic groups that have been separated by more than 30 thousand years will create extensive LD that will persist for 10–20 generations after the admixture event, or longer if the admixture process is more complex and continuous (Briscoe et al. 1994; Pfaff et al. 2001). This LD could provide the basis for identifying chromosomal regions that are specifi-

H. E. Collins-Schramm · B. Chima · D. J. Operario
M. F. Seldin (✉)
Rowe Program in Human Genetics,
Departments of Biological Chemistry and Medicine,
University of California at Davis,
One Shields Avenue, Davis, CA, 95616-8669, USA
Tel.: +1-530-7546017, Fax: +1-530-7546015,
e-mail: mfseldin@ucdavis.edu

L. A. Criswell
Rosalind Russell Medical Research Center for Arthritis,
University of California, San Francisco, CA, 94143, USA

cally associated with disease-promoting or disease-preventing loci in one but not the other ethnic group.

One candidate population for admixture mapping, the African American (AA) population, has several advantageous qualities that make it suitable for this approach. The European American (EA) and African (AF) populations are separated by 50,000 to 100,000 years (Cavalli-Sforza et al. 1994; Underhill et al. 2001). The AA population is the result of relatively recent admixture between these two groups, as most of the AF ancestors of today's African Americans were brought to the United States approximately 300 years ago (Tanner 1995). In addition, the size of this population (approximately 35.5-million individuals) results in large patient bases that make association studies feasible (US Census Bureau, see website). Several diseases, including type 2 diabetes, hypertension, prostate cancer and systemic lupus erythematosus, are thought to have major differences in the genetic risk of disease between AF and EA populations, and therefore may be appropriate for admixture mapping studies in AAs (Gaskin 1999; Molokhia and McKeigue 2000; Farkas et al. 2000). Many markers with confirmed large frequency differences between EA and AF populations have been reported, and hundreds more are suggested by the results of The SNP Consortium (Shriver et al. 1997; Smith et al. 2001; Collins-Schramm et al. 2002a; see the website of The SNP Consortium). Examination has shown that the admixture dynamics of the AA population are likely to be complex and, at least to some degree, continuous (Parra et al. 2001). This type of admixture potentially leads to more extensive LD and therefore more power for admixture mapping. It can also lead to false positives, but it is possible to differentiate the true positives from the false positives with knowledge of individual admixture (Parra et al. 2001). Lastly, our recent studies provide strong evidence that markers with large frequency differences between EAs and one AF population have very limited variation between different AF populations (Collins-Schramm et al. 2002b). This suggests that differences in allele frequencies of such ancestry, informative/ethnic-difference markers within the various sub-populations that were subjected to the slave trade, will not be a major confounding problem.

There have been few empirical studies of the strength and extent of LD in the AA population, and the results of these studies have been partially discrepant. Parra et al. (1998) reported strong LD between FY-null and AT3, markers separated by 22 cM. This result was later confirmed by Lautenberger et al. (2000), who also examined LD between FY-null and other markers in the region, and by Parra et al. (2001). In contrast, others reported similar levels of LD to those seen in non-admixed populations – detectable to a maximum of about 50 kb (Gabriel et al. 2002; Rybicki et al. 2002a). These differences may be due to the difference in markers used by these groups. Notably, the groups observing strong LD used markers selected for relatively large differences between EAs and AFs, while the other studies used unselected markers. However, all of the marker comparisons in both the Parra et al. (2001) and the Lautenberger et al. (2000) reports in-

involved the marker FY-null, a marker with nearly fixed divergent frequencies in the parental populations. There is a large amount of evidence that this allele frequency difference is due to the effects of selection, which could result in increased LD in the parental populations, which were not examined (Hamblin and Di Rienzo 2000). It has yet to be determined whether the extended LD observed is specific to this marker or is a consistent phenomenon in the African American population. In addition, previous studies have not extensively examined unlinked loci. This aspect is critical since insufficient time since admixture or differences in admixture ratios in different AA individuals due to the resulting population substructure could lead to LD between unlinked loci. This LD would lead to false positive results in an admixture mapping study. Pfaff et al. (2001) did a preliminary examination of the extent of LD between unlinked markers. They found that up to a third of 37 unlinked marker-pairs did show significant LD. However, all of their comparisons were based on different pairwise combinations of 11 markers, and all comparisons that were significant included the marker FY.

A recent study of AA LD examined 109 unselected microsatellite markers (Rybicki et al. 2002b). A wide range of LD levels were detected. In addition, no correlation between LD and recombination distance and only a weak negative correlation between LD and physical distance was seen. However, for the subset of these markers for which the EA allele frequencies were available in the CEPH database, the authors found that markers with large allele frequency differences between EA and AA had more extensive LD that correlated better with both recombination and physical distances than markers with small allele frequency differences EA and AA.

Before the input of the resources required for genome-wide disease screens by admixture mapping can be justified, the extent and characteristics of admixture linkage disequilibrium in the African American population need to be examined in more detail. The extent of linkage disequilibrium should be examined in several genomic regions using multiple genetic markers. In addition, a larger scale examination of the strength and frequency of linkage disequilibrium between unlinked markers with large allele frequency differences between the parental populations will be required to determine if false positive results would hinder admixture mapping in this population.

This investigation was undertaken to provide a more detailed examination of the extent of LD created by admixture in AAs. A total of 52 diallelic markers informative for ancestry were utilized in 133 linked and 133 unlinked pairwise LD analyses. These included linked markers from four different chromosomal regions. Short insertion-deletion polymorphisms (indels) and single nucleotide polymorphisms (SNPs) were used, since diallelic markers are likely to be the primary markers utilized in admixture mapping studies (the requirement for large parental allele frequency differences means that many markers must be screened to saturate the genome with appropriate markers, a requirement that will likely only be accomplished with the use of SNPs). The results indicate that, unlike in rep-

representatives of both parental populations, consistent LD extends for genomic intervals of greater than 10 cM and 15 megabases (Mb) in the admixed AA population.

Materials and methods

Populations and samples

Blood- or buccal-cell samples were obtained from all individuals, according to protocols and informed-consent procedures approved by institutional review boards, and were labeled with an anonymous code number. None of the individuals were first-degree relatives of each other, and ethnicities were self-described. The EA individuals were random volunteers from northern California. AF samples were from the Zimbabwe Shona (a Bantu-speaking group) and were supplied by R. Cooper according to National Institutes of Health guidelines. The direct descendants of these individuals were unlikely to have been part of the slave trade, which mainly occurred along the West African coast. However, our previous research has shown that for markers with large standard variances between these Zimbabwean samples and EAs, the variance between the Zimbabwean samples and a collection of samples from Nigeria was very small (Collins-Schramm et al. 2002b). Therefore this population should be appropriate as a representative parental population for admixture mapping studies in today's AAs. For AA individuals, approximately 75% of samples were volunteers from northern California, and 25% were from a wide distribution of other geographic locations in the United States. In the AA samples, all individuals had no known parents or grandparents whom they would describe as being of direct European, African, or Amerindian heritage.

Only those individuals typed for both markers in a LD comparison were utilized, so no missing data calculations were required. Sample size was an average of 259 for AA, 137 for EA and 142 for AF LD analyses. DNA samples were prepared from blood or buccal-cells as previously described (Bali et al. 1999).

Markers and conditions

All markers utilized in this study were biallelic. Forty-eight were short insertion/deletion polymorphisms (MIDs) originally identified by the Marshfield Center for Medical Genetics (see website of the Marshfield Center for Medical Genetics), and four were SNPs selected from the SNP consortium (see website of the SNP Consortium). Chromosome 5 was particularly saturated, with 24 markers, to allow linkage disequilibrium calculations for a large range of distances. Six markers were in a linkage group on chromosome 22, five were in a linkage group on chromosome 6 and four were in a linkage group on chromosome 16. The remaining markers were used for unlinked comparisons; they included three each on chromosomes 2 and 3, two each on chromosomes 1 and 10, and one each on chromosomes 7, 9 and 13. Detailed marker information and a list of all comparisons examined is available at our web site (see the website of the Rowe Program).

The standard variance (also known as the Wahlund variance and shown as the f value or F_{st}) of each marker was calculated using the representative parental populations described above. It is calculated by the following formula, where μ_x is the frequency of allele 1 in population x and μ_y is the frequency of allele 1 in population y :

$$f = (\mu_x - \mu_y)^2 / [4\mu(1 - \mu)], \text{ where } \mu = 1/2(\mu_x + \mu_y)$$

This value is a measure of the ethnic information provided by a marker, and ranges from 0 (non-informative) to 1 (completely informative).

Genomic positions

The approximate Mb position for each marker was determined by use of the Human Genome Browser (J. Kent, University of Cali-

fornia, Santa Cruz), based on the June 2002 human genome draft assembly (see the website of the UCSC Human Genome Project Working Draft). The genetic map positions were based on Marshfield map positions (see website of The Marshfield Center for Medical Genetics). These were interpolated based on the sequence location of genetic markers on this map that were located within short physical distances flanking the physical location of the markers in the sequence assembly.

Examination of linkage disequilibrium

LD analyses were performed using the Arlequin program, which determines population haplotype frequencies by application of the EM algorithm, and then examines significance of LD by a permutation test. For the EM calculations, 10,000 maximum iterations, 100 starting values, and ten bootstrap replicates were used. For the determination of LD significance, 50,000 permutations were performed with 100 initial conditions. The smallest P value that Arlequin calculates is $P=0.00000$.

D' calculations require individual haplotypic information, which was not available in this study. However, estimated D' values were calculated based on the best estimate haplotypes derived by PHASE, which has been shown to often be superior to the EM algorithm for haplotype reconstruction at the individual level (Stephens et al. 2001). For these calculations, 500 permutations were performed with 500 initial starting conditions (note: multiple runs using these same parameters had identical results). The Arlequin program was then used to calculate D' values from the estimated haplotypes. In addition, for a subset of 20 markers, population haplotype frequencies derived from PHASE were used in Arlequin to calculate P values of LD significance. The P values using these PHASE-predicted haplotypes were nearly identical to the P values obtained when the EM algorithm was used.

Results

Long-range LD is present in AA but not representative parental populations over a 21-cM region of chromosome 5

To provide a detailed examination of LD, a region of chromosome 5 from 118 to 139 cM was chosen, in which preliminary examination had identified several markers with high standard variances (f or F_{st} values) between AF and EA populations (Collins-Schramm et al. 2002a, 2002b). We reasoned that these markers which are informative for ancestry would provide the best tool for detecting admixture LD (see Introduction). Additional markers were screened, resulting in a panel of 12 markers with standard variances ranging from 0.15 to 0.69. LD between each possible marker-pair was then evaluated in AA, EA, and AF individuals.

Significant LD was found in the AA population in 60 of the 66 pair-wise analyses including markers separated by greater than 10 cM (Table 1). In contrast, only one pair of markers in the EA population (Table 2) and four pairs of markers in the AF population (Table 3) achieved even marginal levels of significance. For example, MID 1191 and MID 719, separated by 12.4 cM, were in strong linkage disequilibrium (P value of 0.00012) in the AA population, but had P values of 0.721 and 0.135 in the EA and AF populations, respectively. No pairwise calculations showed strong LD (P value <0.001) in the EA or AF populations, but 41 linked marker-pairs had strong LD in the AA population. In fact, 18 of the linked marker-pairs

Table 1 *P* values of pairwise linkage disequilibrium on chromosome 5 in African Americans. The standard variance for each marker is shown as an *f* value. Numbers in italics are the *P* values of LD analyses performed using the Arlequin program

Marker ^a :	1272	883	1848	879	1191	569173 ^b	1937	224652 ^b	1013	768	1102	719
cM ^c :	118	118.8	119.5	120.7	126.7	127.7	128.6	130.3	133.7	135.8	136.1	139.1
Mb ^d :	109.41	110.47	111.29	111.84	115.53	116.81	117.82	119.76	126.38	130.85	131.62	134.03
<i>f</i> value:	0.69	0.42	0.25	0.38	0.28	0.36	0.28	0.15	0.18	0.53	0.66	0.30
1272	–	0.00000	0.00002	0.00000	0.00000	0.00000	0.00002	0.00112	0.07229	0.02539	0.00009	0.0001
883	–	–	0.0002	0.00000	0.01672	0.00012	0.01753	0.04506	0.08036	0.02452	0.00484	0.25027
1848	–	–	–	0.00002	0.00000	0.00000	0.10021	0.15920	0.00026	0.10222	0.02531	0.00983
879	–	–	–	–	0.00000	0.00000	0.00002	0.00403	0.00454	0.00346	0.00059	0.00269
1191	–	–	–	–	–	0.00000	0.00000	0.00002	0.00000	0.00026	0.00000	0.00012
569173 ^b	–	–	–	–	–	–	0.00000	0.00000	0.00028	0.0002	0.00000	0.0051
1937	–	–	–	–	–	–	–	0.00001	0.00155	0.04381	0.00000	0.00508
224652 ^b	–	–	–	–	–	–	–	–	0.00028	0.00028	0.00004	0.00991
1013	–	–	–	–	–	–	–	–	–	0.00006	0.00004	0.02744
768	–	–	–	–	–	–	–	–	–	–	0.00000	0.00002
1102	–	–	–	–	–	–	–	–	–	–	–	0.00002

^aMarkers included Marshfield Indels (designated by MID numbers) and SNPs (see footnote b)

^bTwo SNPs are identified by TSC numbers as designated by the SNP Consortium (see <http://snp.cshl.org/>)

^cGenetic map positions based on Marshfield map positions. These were interpolated based on the sequence location of genetic markers

on this map located within short physical distances flanking the physical location of the markers in the sequence assembly

^dThe approximate megabase position for each marker was determined by use of the Human Genome Browser (J. Kent, University of California, Santa Cruz), based on the June 2002 human genome draft assembly, <http://genome.ucsc.edu/>

Table 2 *P* values of pairwise linkage disequilibrium on chromosome 5 in European Americans. The standard variance for each marker is shown as an *f* value. Numbers in italics are the *P* values of LD analyses performed using the Arlequin program

Marker ^a :	1272	883	1848	879	1191	569173 ^b	1937	224652 ^b	1013	768	1102	719
cM ^c :	118	118.8	119.5	120.7	126.7	127.7	128.6	130.3	133.7	135.8	136.1	139.1
Mb ^d :	109.41	110.47	111.29	111.84	115.53	116.81	117.82	119.76	126.38	130.85	131.62	134.03
<i>f</i> value:	0.69	0.42	0.25	0.38	0.28	0.36	0.28	0.15	0.18	0.53	0.66	0.30
1272	–	0.35506	0.27773	0.59976	0.80889	0.97511	0.59476	0.64568	0.39265	0.01273	0.08253	0.26802
883	–	–	0.94341	0.18739	0.53644	0.91037	0.59075	0.79486	0.10418	0.74818	0.48482	0.5003
1848	–	–	–	0.80249	0.38111	0.94244	0.90585	0.36674	0.58722	0.63516	0.33846	0.35187
879	–	–	–	–	0.50358	0.38633	0.06984	0.77963	0.89549	0.99016	0.32381	0.09804
1191	–	–	–	–	–	0.40658	0.3818	0.88379	0.92339	0.03215	0.77917	0.72088
569173 ^b	–	–	–	–	–	–	0.23954	0.97622	0.41553	0.35585	0.16618	0.42651
1937	–	–	–	–	–	–	–	0.88181	0.13022	0.72991	0.86768	0.22784
224652 ^b	–	–	–	–	–	–	–	–	0.15715	0.79639	0.28803	0.42053
1013	–	–	–	–	–	–	–	–	–	0.99998	0.28668	0.31992
768	–	–	–	–	–	–	–	–	–	–	0.22745	0.63346
1102	–	–	–	–	–	–	–	–	–	–	–	0.34575

^aMarkers included Marshfield Indels (designated by MID numbers) and SNPs (see footnote b)

^bTwo SNPs are identified by TSC numbers as designated by the SNP Consortium (see <http://snp.cshl.org/>)

^cGenetic map positions based on Marshfield map positions. These were interpolated based on the sequence location of genetic markers

on this map located within short physical distances flanking the physical location of the markers in the sequence assembly

^dThe approximate megabase position for each marker was determined by use of the Human Genome Browser (J. Kent, University of California, Santa Cruz), based on the June 2002 human genome draft assembly, <http://genome.ucsc.edu/>

showed the smallest *P* value ($P=0.00000$) that can be observed with the analytic program utilized in this study (see Materials and methods).

Strong LD in AA

is only observed between linked markers

These studies were next extended to examine a total of 133 pairs of linked markers and 133 pairs of unlinked markers in African Americans (Fig. 1). The pairs of linked markers included regions of chromosomes 22, 6 and 16 in

addition to the chromosome 5 region discussed above. The pairs of unlinked markers were selected to closely match the sum of standard variance of the linked pairs of markers. For the group with summed $f > 0.8$, 29 linked marker-pairs with an average f of 0.98 ± 0.13 (SD) and 29 unlinked marker-pairs with an average f of 0.98 ± 0.11 were examined. For the group with summed f between 0.45 and 0.8, 63 linked marker-pairs with an average f of 0.59 ± 0.9 and 63 unlinked marker-pairs with an average f of 0.64 ± 0.11 were examined. For the group with summed f between 0.2 and 0.45, 41 linked marker-pairs with an average f of

Table 3 *P* values of pairwise linkage disequilibrium on chromosome 5 in Africans. The standard variance for each marker is shown as an *f* value. Numbers in italics are the *P* values of LD analyses performed using the Arlequin program

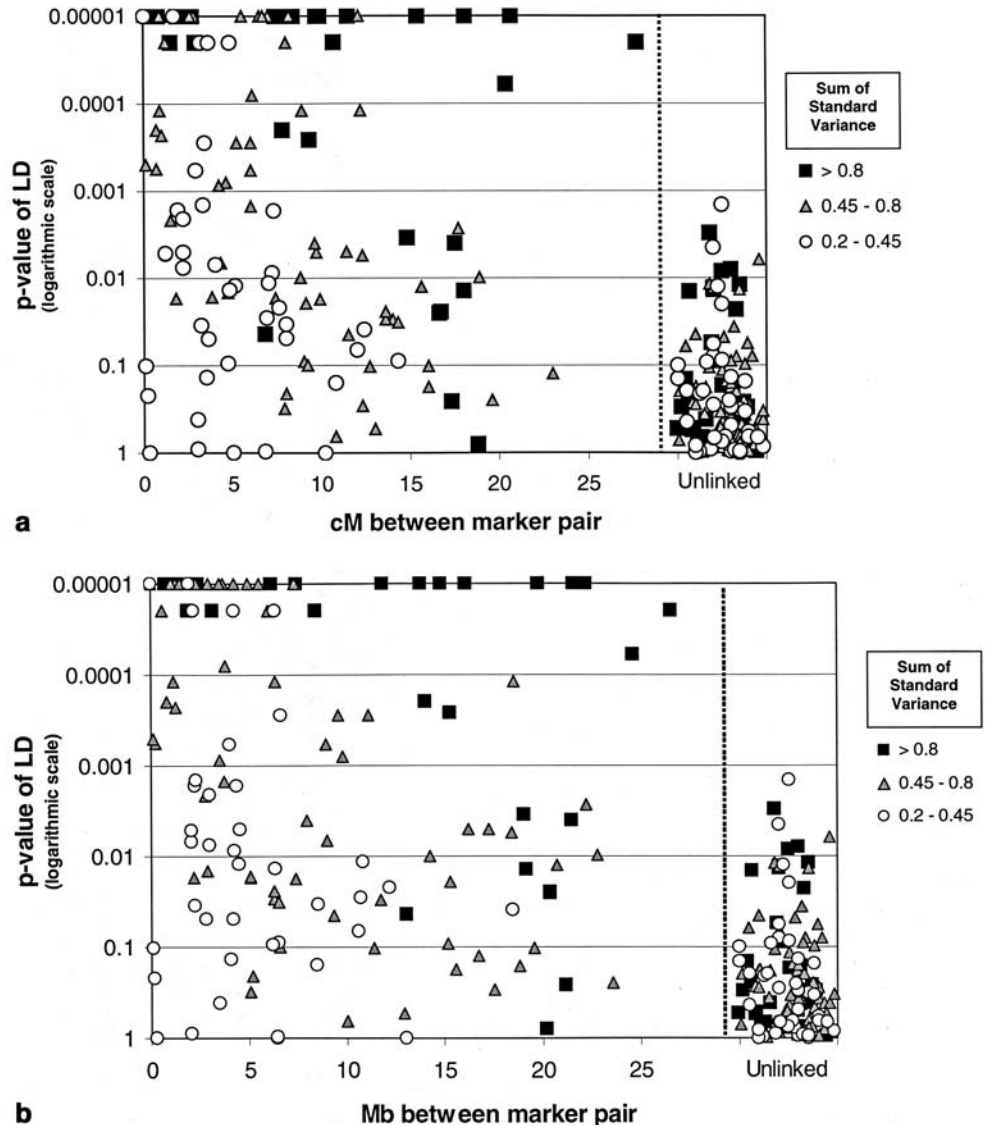
Marker ^a :	1272	883	1848	879	1191	569173 ^b	1937	224652 ^b	1013	768	1102	719
cM ^c :	118	118.8	119.5	120.7	126.7	127.7	128.6	130.3	133.7	135.8	136.1	139.1
Mb ^d :	109.41	110.47	111.29	111.84	115.53	116.81	117.82	119.76	126.38	130.85	131.62	134.03
<i>f</i> value:	0.69	0.42	0.25	0.38	0.28	0.36	0.28	0.15	0.18	0.53	0.66	0.30

1272	–	0.61732	0.73393	0.10946	0.23077	0.05965	0.11958	0.75135	0.04925	0.58132	0.51373	0.73383
883	–	–	0.68185	0.93372	0.00466	0.50176	0.25332	0.69164	0.93632	0.43787	0.73504	0.10483
1848	–	–	–	0.28622	0.95426	0.57818	0.87015	0.25712	0.83749	0.92504	0.08987	0.15269
879	–	–	–	–	0.0219	1	0.96931	0.27227	0.89375	0.21683	0.9987	0.55434
1191	–	–	–	–	–	0.60853	0.42984	0.30577	0.24771	0.26812	0.3649	0.13525
569173 ^b	–	–	–	–	–	–	0.59209	0.89112	0.75404	0.11791	0.97154	0.68995
1937	–	–	–	–	–	–	–	0.3074	0.03187	0.36375	0.48367	0.80845
224652 ^b	–	–	–	–	–	–	–	–	0.28056	0.93901	0.66131	0.13666
1013	–	–	–	–	–	–	–	–	–	0.84781	0.18641	0.05551
768	–	–	–	–	–	–	–	–	–	–	0.0029	0.09495
1102	–	–	–	–	–	–	–	–	–	–	–	0.39711

^aMarkers included Marshfield Indels (designated by MID numbers) and SNPs (see footnote b)
^bTwo SNPs are identified by TSC numbers as designated by the SNP Consortium (see <http://snp.cshl.org/>)
^cGenetic map positions based on Marshfield map positions. These were interpolated based on the sequence location of genetic mark-

ers on this map located within short physical distances flanking the physical location of the markers in the sequence assembly
^dThe approximate megabase position for each marker was determined by use of the Human Genome Browser (J. Kent, University of California, Santa Cruz), based on the June 2002 human genome draft assembly, <http://genome.ucsc.edu/>

Fig. 1a, b Decay of linkage disequilibrium in the African American population as a function of distance. Each point on the graph represents a pairwise LD analysis, ordered along the abscissa in panel **a** by the cM-distance between the marker-pair and in panel **b** by the Mb-distance between the marker-pair. In both panels, *P* values of LD between the marker-pair are plotted on the ordinate, on a logarithmic scale ranging from 1 to 0.00001. Marker-pairs with the minimum *P* value obtainable by the Arlequin program used for analysis (0.00000) were plotted as 0.00001. Each marker-pair is grouped into a category by the sum of the standard variances (*f*) of its two markers. For comparison, on the right of each graph are the *P* values of LD between unlinked markers matched for summed *f* to the linked comparisons (these are spaced along the abscissa randomly to facilitate viewing)



0.32±0.06 and 41 unlinked marker-pairs with an average f of 0.32±0.06 were examined.

There was a clear distinction between the strength of LD observed in the linked compared to unlinked marker-pairs. Evidence for strong LD ($P<0.001$) was observed for 56 of the 133 pairs of linked markers, including ten marker-pairs separated by greater than 10 cM. In contrast, strong LD ($P<0.001$) was not reached for any of the 133 unlinked marker-pairs (Fig. 1).

Nominally significant LD ($P<0.05$) was observed in a total of 104 pairs of the 133 linked marker-pairs examined (Fig. 1 and website of the Rowe Program). However, nominal evidence for LD was also observed in the analysis of unlinked markers (19 of 133 pairs with P values <0.05). A subset of 36 unlinked pairs was also examined in the AF and EA populations. None of these reached significance ($P<0.05$) in AF or EA. These marker-pairs included 11 of the 19 pairs of unlinked markers that were significant in the AA analysis (see Electronic database information, Rowe Program, Web Table of LD results). Thus, the low levels of LD between unlinked marker-pairs observed in the AA population were not present in representatives of either parental population.

LD in the AA population

persists for extensive cM and Mb distances

The percent of linked comparisons with evidence of strong LD, as well as the overall strength of LD, as measured by the P values, decreased as a function of distance (Fig. 1, Table 4). This decrease appeared to be slightly more consistent with the physical distance (Mb) (Fig. 1b) separating the marker-pairs than the interpolated genetic map positions (Fig. 1a). However, there were several marker-pairs with evidence for strong LD at distances greater than 10 cM and 15 Mb. For example, MID 1683 and MID 879 were in strong LD with a P value of 0.00001, despite being separated by 20.7 cM and 21.6 Mb. MID 106 and MID 107 also achieved a P value of 0.00001; they are separated by 12.1 cM, but by only 3.5 Mb.

For another measure of the strength of LD in the AA population, D' values were calculated for each of the

marker-pairs showing strong evidence for LD by the permutation analysis ($P<0.001$). D' was calculated from the predicted individual haplotypes determined using the PHASE program (see Materials and methods). All marker-pairs produced D' values of 0.3 or higher and many had D' values of greater than 0.6 (Fig. 2). The strongest D' values were found between markers separated by less than 5 Mb; however, D' values of greater than 0.5 were observed for several marker pairs separated by greater than 10 Mb.

Correlation of AA LD

with the standard variance of markers

LD decayed differently for marker-pairs with large summed standard variances than those with small standard variances (Fig. 1, Table 4). For marker-pairs separated by many Mb, a larger percentage of those with summed $f>0.8$ were in significant LD than those with lower summed f values (Table 4). To quantify this difference, the least square regression of the negative log of P values on the cM-distance separating markers was calculated for markers sets with different standard variances, setting the y intercept at 5 (a P value of 0.00001). For markers with summed $f>0.8$, the best-fit slope was -0.098 (95% CI = -0.140 to -0.056). In contrast, the best-fit slope for markers with summed f between 0.45 and 0.8 was -0.236 (95% CI = -0.272 to -0.201) and the best-fit slope for markers with summed f between 0.2 and 0.45 was -0.462 (95% CI = -0.569 to -0.354). Figure 1b shows the same groups plotted by Mb separation between markers rather than cM separation. The graphs are obviously very similar, and show the same difference in LD decay dependent on summed f value, with a best fit slope of -0.086 (95% CI = -0.120 to -0.052) for comparisons with summed $f>0.8$ and a best-fit slope of -0.412 (95% CI = -0.518 to -0.306) for comparisons with summed f between 0.2 and 0.45. However, in both graphs, the summed f value does not appear to completely determine the slope of LD decay, as creating more refined groupings based on summed f value does not reduce the scatter in data points (data not shown). Analyses of these data, using either the product of f or the prod-

Table 4 Comparison of long-range LD in marker pairs grouped by standard variation

Interval (Mb) ^a	$f>0.8^b$			$0.45<f<0.8$			$f<0.45$		
	Pairs ^c	$P<0.001^d$	Percent ^e	Pairs	$P<0.001$	Percent	Pairs	$P<0.001$	Percent
<2.5	6	6	100%	10	10	100%	13	3	23%
2.5–5.0	1	1	100%	11	8	73%	13	2	15%
5.0–10.0	3	3	100%	19	7	37%	9	2	22%
10.0–15.0	5	4	80%	5	1	20%	5	0	0%
15.0–20.0	5	3	60%	11	1	9%	1	0	0%
20.0–25.0	9	4	44%	3	0	0%	0	n.d.	n.d.

^aLength of intervals between marker-pairs analyzed for LD

^bSum of the f values for the marker-pairs analyzed

^cNumber of marker-pairs examined

^dNumber of marker-pairs with evidence for strong LD ($P<0.001$)

^ePercent of marker-pairs examined with highly significant LD ($P<0.001$)

Fig. 2 Estimated D' values for marker-pairs with strongly significant permutation results. The 55 marker-pairs with LD P values of ≤ 0.001 by permutation analysis are plotted by Mb-distance between the two markers and estimated D' values. Estimated D' values were calculated from individual haplotypes constructed from genotypic data by the PHASE program

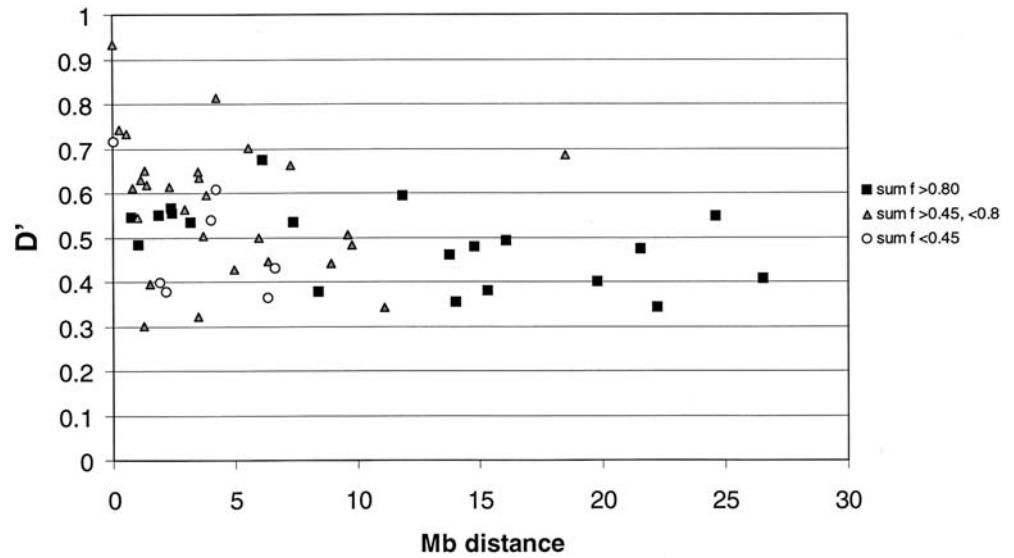
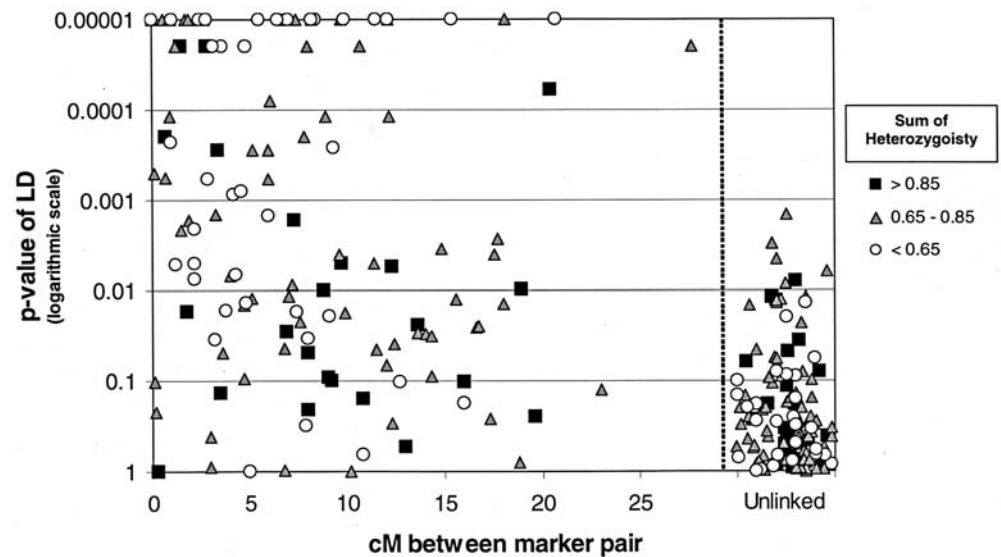


Fig. 3 Decay of linkage disequilibrium in the African American population is independent of marker heterozygosity. All marker-pairs shown in Fig. 1a were grouped into categories based on the summed heterozygosity of the two markers rather than summed standard variance (f value)



uct of δ (allele frequency difference), showed a similar pattern of results (data not shown). The least square regression favored the summed f or product of f rather than the product of δ values consistent with predicted information content (McKeigue 1998).

An alternative possibility is that the heterozygosity of markers, rather than the summed standard variance, determines the slope of LD decay. To examine this possibility, Fig. 3 plots the LD analyses grouped by summed heterozygosity rather than summed standard variance. There is no obvious difference between points in the three groups, and the best-fit slopes are not significantly different between comparisons with summed heterozygosities greater than 0.85 (95% CI = -0.34 to -0.17) and comparisons with summed heterozygosities less than 0.65 (95% CI = 0.22 to -0.09).

Discussion

Our examination of 133 linked marker comparisons from several genomic regions clearly demonstrates that LD in the AA population, unlike representative parental EA or AF populations, can extend for many cM. LD was detectable by both permutation testing and D' calculations from haplotypes determined by both the EM algorithm and by PHASE. These results, in conjunction with the previous work of Lautenberger et al. (2000) examining the Duffy region, strongly suggest that the phenomenon of extensive LD in the AA population is genome-wide.

The AA population also had an increase in the percentage of unlinked comparisons that showed evidence of LD, although the strength of this LD was orders of magnitude less than that seen between many linked comparisons. This increase in nominal LD between unlinked markers is likely due to varying admixture ratios between AA indi-

viduals leading to population substructure, possibly as a result of non-random mating. An additional explanation, as suggested by Pfaff et al. (2001), is that the increase may be the result of continuous gene flow into the AA population. In this hypothesis, one or both parental populations contribute to the admixed population over multiple generations, resulting in a complex genetic history of present day admixed individuals and effectively creating a population that has not had enough time for recombination to reduce the LD between unlinked sites. The existence of strong linkage disequilibrium over distances of 10 cM also supports this model of admixture for the AA population. However, as seen in Fig. 1 and Fig. 2, there were no P values between unlinked marker comparisons that reached the strong levels of LD ($P < 0.001$) that many linked comparisons attained. Thus, although the likelihood of LD between unlinked markers must be appreciated in an admixture mapping study of the AA population, it should be relatively easy to choose a level of significance that would eliminate these false-positive signals.

The results of Fig. 1 and Table 4 strongly suggest that the strength of LD between linked markers in AA is dependent not only on the distance between the markers but also on their standard variance. This result is in agreement with recent studies using microsatellite markers (Rybicki et al. 2002b), and is further evidence that the extended LD seen in the AA population is due to admixture. The dependence of LD strength on standard variance is the likely explanation of discrepancies in the extent of LD observed by previous researchers. Studies done with unselected markers will have few marker-pairs with large summed standard variances, and therefore only rare cases of extended LD will be seen. This result emphasizes the requirement for a genomic set of markers with large standard variances between EAs and AFs, as suggested by several other investigators (Dean et al. 1994; Smith et al. 2001).

Standard variance and cM-distance, however, can not fully explain the strength of LD observed, as seen by the large scatter still present in each data group in Fig. 1. For example, MIDs 1191 and 768, separated by 9.3 cM with a summed f of 0.81, produced a P value of 0.0003, while MIDs 1937 and 768, also with a summed f of 0.81 but only separated by 6.8 cM, produced a P value of 0.044. These inconsistencies indicate that other variables may also be present. Figure 3 demonstrates that the heterozygosity of the markers is not likely to be one of these variables. One likely variable is differences in recombination frequency not reflected by the genetic map that relied on CEPH European and European-American family data. In the current study, the decay of LD appeared to marginally correspond better with physical distance, although this observation is based on only a limited number of markers (e.g., MID 106 and MID 107, see above). Other variables that might affect the strength of LD include marker stability, structure within one of the founding populations and preexisting LD within one of the founding populations. However, available data do not provide any inference that these possibilities are major factors: (1) marker stability must be high in one or both founding populations to result

in the observed large standard variance between founding populations; (2) intra-ethnic variation for markers with high standard variance is limited (Collins-Schramm et al. 2002b); (3) little LD was observed between markers within each of the representatives of the putative populations (Table 1 and unpublished observations).

Methods to examine allelic association between markers and disease status are only one method of performing admixture mapping. Potentially more powerful methods of performing admixture mapping have been proposed and are currently under development (McKeigue et al. 2000; Hongzhe Li, unpublished method). These methods vary in their approach to analysis (e.g., some use Bayesian, while others use maximum likelihood statistical approaches); however, they are based on the same general concept. The idea is that one can maximize power in admixture mapping by not examining allelic association per se, but by rather attempting to determine the ancestral identity of each chromosomal region. For example, imagine a disease susceptibility allele that was contributed to the AA population primarily from the AF population. In this case, one would look for genomic regions where admixed individuals with disease showed an increase in the inheritance of DNA from the AF population that was statistically different from that seen in the rest of the genome or in control AAs. If markers were completely fixed for different alleles in the two parental populations, then this would be the same as examining allelic association, but even nearly fixed divergent markers are so rare it would be impossible to saturate a genome with them. Instead, these methods of analysis increase the ability to assign a chromosomal region to a given ancestry by utilizing information from several markers located at distances much smaller than the average phase transition of ancestry. These multilocus approaches have the potential to be much more powerful than simply examining allelic association. Because of this, examination of allelic LD created by admixture, such as we have done here, demonstrates the feasibility of admixture methods but may not reflect the true strength of admixture association that could be detectable by multilocus admixture mapping methods.

Acknowledgements Support for this research was provided by National Institutes of Health grants AR44804 and DK57249. The authors thank Yolanda Figueroa for assistance in application of computational algorithms and programs and Richard Cooper for providing anonymous African DNA samples.

Electronic database information

URLs for data in this article are as follows:

- US Census Bureau, <http://www.census.gov> (for size of the AA population)
- The SNP Consortium, <http://snp.cshl.org/> (for initial screening information of the SNPs used in this study and other potential SNPs for admixture mapping)
- The Marshfield Center for Medical Genetics, <http://research.marshfieldclinic.org/genetics> (for initial screening information of the MIDs utilized in this study, in-

cluding allele frequencies in several populations, and for cM positions)

- UCSC Human Genome Project Working Draft, <http://genome.ucsc.edu/> (for Mb positions of MIDs)
- UC Davis, Rowe Program, http://roweprogram.ucdavis.edu/Seldin_lab_confidential/Linkage_Disequilibrium.pdf (for Web Table of all pairwise LD analyses in AA population)

References

- Bali D, Gourley IS, Kostyu DD, Goel N, Bruce I, Bell A, Walker DJ, Tran K, Zhu DK, Costello TJ, Amos CI, Seldin MF (1999) Genetic analysis of multiplex rheumatoid arthritis families. *Genes Immun* 1:28–36
- Briscoe D, Stephens JC, O'Brien SJ (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. *J Hered* 85:59–63
- Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, New Jersey
- Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, Cooper R, Li H, Seldin MF (2002a) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am J Hum Genet* 70:737–750
- Collins-Schramm HE, Kittles RA, Operario DJ, Weber JL, Criswell LA, Cooper RS, Seldin MF (2002b) Markers that discriminate between European and African ancestry show limited variation within Africa. *Hum Genet* 111:566–569
- Dean M, Stephens JC, Winkler C, Lomb DA, Ramsburg M, Boaze R, Stewart C, Charbonneau L, Goldman D, Albaugh BJ, Goedert JJ, Beasley RP, Hwang L, Buchbinder S, Weedon M, Johnson PA, Eichelberger M, O'Brien SJ (1994) Polymorphic admixture typing in human ethnic populations. *Am J Hum Genet* 55:788–808
- Farkas A, Marcella S, Rhoads GG (2000) Ethnic and racial differences in prostate cancer incidence and mortality. *Ethn Dis* 10:69–75
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. *Science* 296:2225–2229
- Gaskin R (1999) Diet, diabetes, hypertension and blacks. *Ethn Dis* 9:272–277
- 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
- Jorde LB, Watkins WS, Kere J, Nyman D, Eriksson AW (2000) Gene mapping in isolated populations: new roles for old friends? *Hum Hered* 50:57–65
- 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
- Lonjou C, Collins A, Morton NE (1999) Allelic association between marker loci. *Proc Natl Acad Sci USA* 96:1621–1626
- 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
- McKeigue PM, Carpenter JR, Parra EJ, Shriver MD (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann Hum Genet* 64:171–186
- Mohlke KL, Lange EM, Valle TT, Ghosh S, Magnuson VL, Silander K, Watanabe RM, Chines PS, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2001) Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. *Genome Res* 11:1221–1226
- Molokhia M and McKeigue P (2000) Risk for rheumatic disease in relation to ethnicity and admixture. *Arthritis Res* 2:115–125
- 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
- Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, Bonilla C, Sylvester N, Parrish-Gause D, Garvey WT, Jin L, McKeigue PM, Kamboh MI, Ferrell RE, Pollitzer WS, Shriver MD (2001) Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *Am J Phys Anthro* 114:18–29
- Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, Shriver MD (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am J Hum Genet* 68:198–207
- Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517
- Rybicki BA, Iyengar SK, Harris T, Liptak R, Elston RC, Maliairik MJ, Iannuzzi MC (2002a) Prospects of admixture linkage disequilibrium mapping in the African-American genome. *Cytometry* 47:63–65
- Rybicki BA, Iyengar SK, Harris T, Liptak R, Elston RC, Sheffer R, Chen KM, Major M, Maliairik MJ, Iannuzzi MC (2002b) The distribution of long range admixture linkage disequilibrium in an African-American population. *Hum Hered* 53:187–196
- 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
- Smith MW, Lautenberger JA, Doo Shin H, Chretien J, Shrestha S, Gilbert DA, O'Brien SJ (2001) Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. *Am J Hum Genet* 69:1080–1094
- Stephens M, Smith MJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–989
- Tanner HH (1995) The settling of North America. MacMillan, New York
- Terwilliger JD and Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotech* 9:578–594
- Tsui LC, Buetow K, Buchwald M (1986) Genetic analysis of cystic fibrosis using linked DNA markers. *Am J Hum Genet* 39:720–728
- Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65:43–62