European Journal of Human Genetics (2005) 13, 677–686
© 2005 Nature Publishing Group All rights reserved 1018-4813/05 $30.00
www.nature.com/ejhg

ARTICLE

Linkage disequilibrium patterns vary substantially among populations

Sarah L Sawyer[1,3], Namita Mukherjee[2,4], Andrew J Pakstis[2], Lars Feuk[1], Judith R Kidd[2], Anthony J Brookes[1,5,6] and Kenneth K Kidd*[,2,6]

[1]*Center for Genomics and Bioinformatics, Karolinska Institute, Berzelius väg 35, Stockholm 171 77, Sweden;*
[2]*Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520-8005, USA*

A major initiative to create a global human haplotype map has recently been launched as a tool to improve the efficiency of disease gene mapping. The 'HapMap' project will study common variants in depth in four (and to a lesser degree in up to 12) populations to catalogue haplotypes that are expected to be common to all populations. A hope of the 'HapMap' project is that much of the genome occurs in regions of limited diversity such that only a few of the SNPs in each region will capture the diversity and be relevant around the world. In order to explore the implications of studying only a limited number of populations, we have analyzed linkage disequilibrium (LD) patterns of three 175–320 kb genomic regions in 16 diverse populations with an emphasis on African and European populations. Analyses of these three genomic regions provide empiric demonstration of marked differences in frequencies of the same few haplotypes, resulting in differences in the amount of LD and very different sets of haplotype frequencies. These results highlight the distinction between the statistical concept of LD and the biological reality of haplotypes and their frequencies. The significant quantitative and qualitative variation in LD among populations, even for populations within a geographic region, emphasizes the importance of studying diverse populations in the HapMap project to assure broad applicability of the results.

European Journal of Human Genetics (2005) **13,** 677–686. doi:10.1038/sj.ejhg.5201368
Published online 19 January 2005

Keywords: haplotype map; SNPs; population genetics; linkage disequilibrium; genetic diversity

Introduction

Nonrandom association of alleles, commonly referred to as linkage disequilibrium (LD), has become an important aspect of studies on population structure and human evolution.[1–3] The observation that LD is frequently seen for close markers in non-African populations but less so in sub-Saharan populations is one of the strong lines of evidence for the African origin of all modern humans.[4–9] LD is also considered to be especially valuable in studies to map genes determining susceptibility for common complex disorders.[1,10–13] As they have very low mutation rates and are numerous, single-nucleotide polymorphisms (SNPs) are the focus for defining nonrandom association between the different allelic forms. The first of the larger-scale LD studies have found evidence for the existence of finite LD blocks (or haplotype blocks), regions with limited numbers of haplotypes, presumably due in part to limited ancestral recombination in these regions. These blocks appear to be separated by regions that tend to randomize the flanking blocks.[14] In the HLA region, the blocks have

*Correspondence: Professor KK Kidd, Department of Genetics, Yale University School of Medicine, 333 Cedar Street, PO Box 208005, New Haven, CT 06520-8005, USA. Tel.: +1 203 785 2654; Fax: +1 203 785 6568; E-mail: kenneth.kidd@yale.edu

[3]Current address: Department of Medical Genetics, University of Calgary, 3330 Hospital Dr NW, Calgary, Alberta, Canada T2N 4N1.

[4]Current address: TCG-ISI Centre for Population Genomics, Building B, Bengal Intelligent Park Ltd, Kolkata, India.

[5]Current address: Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK.

[6]These authors contributed equally to this work.

Received 2 March 2004; revised 1 December 2004; accepted 8 December 2004

been shown to be defined by recombination hot spots.[15–17] Based on these observations, the 'HapMap' project has been launched to study the pattern of LD blocks in the human genome among a few populations that are being chosen to represent global human diversity.[18,19] However, there are fundamental issues still unanswered. Recent analyses argue that random genetic drift in finite populations will generate regions of high LD and limited haplotypic diversity in regions of uniform recombination if the recombination rate is sufficiently low.[20,21] Thus, while some long 'blocks' exist and probably reflect regions of unusually infrequent recombination, it is possible that much of the genome shows shorter 'blocks' of largely stochastic origin. In those regions, there may be very different patterns of LD in different populations because of their different recent histories and demographies.

Three aspects of LD in humans are well documented: (1) LD varies along the chromosomes with regions of high LD interspersed with regions with little LD,[1,12,22–28] (2) the extent of LD can vary dramatically among populations,[8,29–31] and (3) various different haplotype structures can be reflected as a single LD pattern. Factors responsible for different patterns of LD will likely vary among populations. In accord with this, several specific factors will be responsible for the patterns of LD but relative contributions will likely vary among populations. Earlier studies on various loci have shown that the extent of this variability in haplotype frequencies and LD among populations can be large when multiple diverse populations are examined.[4–6,8,32] Those studies of LD in large numbers of populations have generally involved only short genomic regions. Studies of longer regions (>100 kb) have generally involved only a few populations.[12,20,22,23,26,33] With these considerations in mind, we have designed a study of LD in three regions of 175 to 300 kb, each with 20–22 SNPs, in 16 different populations to explore the differences in the pattern of LD among populations. These 16 populations generally focus on African and European diversity, but include populations from most of the geographical spread of modern humans, and thus provide a detailed preliminary overview of how patterns of LD along the chromosome and haplotypes are likely to be distributed globally.

Rather than examining LD as pairwise measures between markers, we have chosen to emphasize the statistic ξ, based on a permutation test, to quantify the deviation from random combinations of the alleles along the chromosomes of a population.[34] One value of this and a closely related statistic[35] is freedom from the restriction to pairwise analysis of biallelic markers. The values of ξ are highly correlated with statistical significance and, when applied pairwise, with Δ^2, also known as r^2 (unpublished results and Supplementary material). The approach used for these data has been to measure LD as a moving window along the chromosome. This single value per segment per population also facilitates comparisons among multiple populations. However, to compensate partially for variation in heterozygosity and to integrate nonrandomness shown by nonadjacent markers, the permutations preserved the observed configurations of the two markers on either side of each intermarker segment.[8,34] This gives us a running measure along the chromosome of the quantitative deviations from the randomness of alleles on either side of the focal segment.

Materials and methods

Markers typed

We have typed 63 SNPs distributed across three genomic regions on two different chromosomes. The list of markers with the UIDs for various databases are given in Table 1. Descriptions of the markers, their allele frequencies, and the typing method used are present in ALFRED (http://alfred.med.yale.edu).[36] Markers were typed variously by RFLP, fluorescence polarization,[37] and dynamic allele-specific hybridization (DASH),[38] and TaqMan.[39] Several markers at *RET-D10S94* and *DRD2-NCAM1* were identified by resequencing using an ethnically diverse panel of 10 individuals (three Africans (Biaka, Lisongo, Yoruba), three Europeans (Adygei, Russian, Dane), two Chinese, and two Native Americans (Pima, Cheyenne)); a potential polymorphism was pursued if a variant occurred at least twice in the 20 chromosomes. Other markers in these two loci were identified from the various SNP databases, published reports, or *in silico* mismatches. All the markers at *TNFRSF6* were identified from HGVbase (http://hgvbase.cgb.ki.se).[40]

Populations studied

We have studied 16 populations focusing on sub-Saharan African populations and European populations with a representation of populations from eastern Asia, the Pacific, North, and South America. Sample sizes and origins of the individual population samples are given in Supplementary Material Table S1 along with the ALFRED UIDs for more detailed descriptions. Population samples averaged approximately 50 individuals, ranging from 23 Nasioi to 116 Irish. Descriptions of the populations can also be found in ALFRED from the allele frequency tables under the site UIDs given in Table 1.

Statistical analysis

Allele frequencies were calculated by gene counting assuming codominant inheritance. Heterozygosities were estimated as $1-\Sigma\bar{p}_i^2$, where $\bar{p}_i$ is the individual allele frequency. F_{ST} was calculated as $\sigma^2/(\bar{p}(1-\bar{p}))$ (Wright[41]). Haplotypes were estimated using the HAPLO program.[42] The haplotype frequencies were used to estimate the LD between any two pair of alleles. LD was calculated as D',[43] Δ^2 (Devlin and Risch[44]), and ξ using HAPLO/P.[8,34] ξ is estimated by permuting the data and determining the

Table 1 Description of 63 polymorphic markers in the three genetic regions studied

Markers in chromosome order	*Distance to next SNP in base pairs*	*Heterozygosity in 16 populations*			*F_{st} 16 populations*	*ALFRED UID for polymorphism*	*db SNP rsID*	*Mode of identifying*
		Average	*Min*	*Max*				
TNFRSF6								
S1	31 532	0.404	0.073	0.500	0.19	SI000771P	rs1837836	SNP database
S2	5497	0.428	0.190	0.500	0.06	SI000772Q	rs1968028	SNP database
S3	7734	0.430	0.180	0.497	0.07	SI000773R	rs717576	SNP database
S4	7807	0.355	0.064	0.500	0.10	SI000774S	rs1561742	SNP database
S5	33 696	0.391	0.064	0.498	0.10	SI000775T	rs1530281	SNP database
S6	31 501	0.372	0.064	0.500	0.10	SI000776U	rs1441734	SNP database
S7	3042	0.336	0.033	0.500	0.14	SI000777V	rs978522	SNP database
S8	2378	0.337	0.039	0.493	0.19	SI000778W	rs4064	SNP database
S9	4029	0.342	0.033	0.500	0.18	SI000779X	rs1926196	SNP database
S10	1905	0.370	0.091	0.500	0.13	SI000780P	rs1571011	SNP database
S11	7790	0.333	0.213	0.499	0.07	SI000781Q	rs1159120	SNP database
S12	3370	0.080	0.000	0.489	0.51	SI000782R	rs2229521	SNP database
S13	3920	0.326	0.020	0.490	0.14	SI000783S	rs1571020	SNP database
S14	519	0.045	0.000	0.259	0.09	SI000784T	rs1051070	SNP database
S15	1820	0.325	0.190	0.499	0.14	SI000785U	rs1468063	SNP database
S16	1856	0.273	0.020	0.444	0.16	SI000786V	rs1926188	SNP database
S17	897	0.381	0.120	0.500	0.14	SI000787W	rs913043	SNP database
S18	8093	0.329	0.190	0.499	0.14	SI000788X	rs874145	SNP database
S19	4828	0.392	0.191	0.499	0.15	SI000789Y	rs1926202	SNP database
S20	13 762	0.327	0.020	0.495	0.09	SI000790Q	rs1571016	SNP database
S21		0.096	0.000	0.483	0.17	SI000791R	rs1319097	SNP database
RET-D10S94								
Intron1 G/C	14 163	0.454	0.252	0.500	0.08	SI000753P	rs2506007	Resequencing
Exon2 *Hae*III	14 100	0.251	0.000	0.497	0.14	SI000194O	rs1800858	Published data
Exon11 *Ban*I	3700	0.237	0.016	0.497	0.22	SI000792S	rs1799939	in silico mismatch
Exon13 *Taq*I	400	0.360	0.194	0.498	0.08	SI000160H	rs1800861	Published data
Intron14 *Sty*I	1404	0.311	0.169	0.499	0.10	SI000793T	rs2075910	Published data
Exon15 *Rsa*IA	7300	0.243	0.056	0.499	0.21	SI000161I	rs1800863	Published data
Intron19 *Taq*I	29 600	0.237	0.037	0.495	0.15	SI000695U	rs2565200	in silico mismatch
3′ *Bse*RI	30 181	0.433	0.125	0.499	0.10	SI000696V	rs2914985	Resequencing
Rsq 8 G/T	10 444	0.219	0.000	0.497	0.18	SI000872R	rs2505553	Resequencing
*Dde*I	25 023	0.223	0.000	0.500	0.16	SI000743O	rs1879317	Resequencing
Rsq22 A/C	15 635	0.305	0.073	0.500	0.15	SI000754Q	rs4987091	Resequencing
*Rsa*I-B	19 968	0.334	0.000	0.468	0.08	SI000869X	rs1915144	dbSNP database
*Sty*I	537	0.224	0.033	0.401	0.06	SI000744P	rs4987092	Resequencing
*Msp*I	5521	0.303	0.000	0.480	0.10	SI000728R	rs4987093	Resequencing
*Sma*I	10 995	0.430	0.127	0.500	0.11	SI000729S	rs4987094	Resequencing
D10S94 *Taq*I	8019	0.310	0.151	0.500	0.09	SI000689X	rs1879316	Resequencing
T indel	25 200	0.343	0.022	0.498	0.14	SI000756S	rs4648315	Resequencing
*Cfo*I	11 284	0.284	0.012	0.500	0.14	SI000690P	rs2266066	Resequencing
*Pst*I	43 956	0.272	0.000	0.500	0.15	SI000691Q	rs2460538	Resequencing
G/A SNP	7098	0.227	0.043	0.375	0.18	SI000755R	rs4648316	Resequencing
*Pvu*II	30 932	0.375	0.122	0.495	0.25	SI000196Q	rs4245604	Resequencing
Intron1 *Msp*I		0.248	0.000	0.494	0.13	SI000195P	rs4646975	Resequencing
DRD2-NCAM1								
5′ *Mae*III	14 826	0.242	0.113	0.465	0.06	SI000797X	rs1799978	Resequencing
Intron1 C/T	18 146	0.335	0.132	0.500	0.07	SI000757T	rs4648317	Resequencing
Intron 1 *Bsl*I	17 126	0.455	0.244	0.500	0.08	SI000758U	rs4648318	Resequencing
Intron 1 *Taq*I 'B'	11	0.277	0.000	0.456	0.26	SI000148N	rs1079597	Published data
Intron 2 *Bcl*I	4547	0.228	0.000	0.494	0.30	SI000140F	rs1079598	in silico mismatch
Intron 2 *Mbo*I	159	0.274	0.022	0.500	0.25	SI000146L	rs2734834	Published data
Intron 2 *Taq*I 'D'	10 546	0.287	0.037	0.498	0.23	SI000139N	rs1800498	Published data
Exon 8 C_326647_1_	2598	0.426	0.256	0.500	0.14	SI000136K	rs6279	Published data
3′ *Hinc*II	7675	0.247	0.000	0.438	0.07	SI000143I	rs2234689	Published data

Table 1 *(Continued)*

Markers in chromosome order	*Distance to next SNP in base pairs*	*Heterozygosity in 16 populations*			*F_{st} 16 populations*	*ALFRED UID for polymorphism*	*db SNP rsID*	*Mode of identifying*
		Average	*Min*	*Max*				
3′ *Taq*I 'A'	54 951	0.374	0.200	0.499	0.15	SI000144J	rs1800497	Published data
Rsq3 C920T	18 706	0.198	0.000	0.495	0.27	SI000762P	rs4987094	Resequencing
NCAM1 T618A	19 734	0.378	0.235	0.498	0.08	SI000710I	rs635358	Resequencing
Rsq4 *Sty*I	21 793	0.322	0.020	0.500	0.08	SI000761O	rs4648321	Resequencing
*Dde*I	27 766	0.357	0.060	0.487	0.28	SI000709Q	rs592903	Resequencing
Intron 14 *Spe*I	22 692	0.189	0.024	0.369	0.05	SI000708P	rs2186798	Resequencing
Intron13 G/T	5649	0.301	0.142	0.449	0.05	SI000794U	rs586903	SNP database
Intron 12 *Bam*HI	20 167	0.379	0.090	0.485	0.07	SI000707O	rs4646982	Resequencing
Intron 8 C/G	5150	0.446	0.231	0.500	0.11	SI000795V	rs686050	SNP database
Intron 7 *Rsa*I	23 126	0.409	0.043	0.499	0.11	SI000704L	rs1545086	Resequencing
Intron 1 *Hha*I		0.258	0.073	0.462	0.06	SI000769N	rs4648322	Resequencing

The 63 SNPs studied with intermarker distances, heterozygosity (average, minimum, and maximum in 16 populations), F_{st} value based on 16 populations, ascertainment method, and database links for each polymorphism (ALFRED UID, dbSNP rs identifier). The allele frequencies in all 16 population samples studied are in ALFRED under the UIDs given. The precise molecular definitions are given in dbSNP under the reference SNP numbers given.

likelihood ratio statistic for the best haplotype frequency estimates for each permuted data set. The value for the true data set, t, is standardized by the mean, μ, and standard deviation, σ, of the likelihood ratio statistic for a large number of such permutations (generally 1000). Finally, this is further standardized by $(2v)^{1/2}/n$, where v is the degrees of freedom of the haplotype system and n is the sample size:

$$\xi = (2v)^{1/2}((t-\mu)/\sigma)/n$$

For the segment test using two (or more) markers on either side of the segment, the permutations are carried out only between the two sets of markers, preserving the configurations of the sites on either side. The advantages of the 2×2 segment test are that it compensates for low heterozygosity at any one SNP and integrates LD contributed by some nonadjacent markers. This approach is analogous to the interval LD approach for generating LD maps[21,45] and has similarities to the moving window approach using entropy.[35] Statistical significance of differences in haplotype frequencies was tested using a likelihood ratio χ^2 heterogeneity test.[46]

Results

Allele frequencies, heterozygosities, and F_{ST}

For most markers typing was complete for at least 95% of individuals; no marker was less than 90% complete. Allele frequencies and numbers of individuals typed for all of the sites in all of the populations can be found in ALFRED[36] under the site UIDs (Table 1). All of the markers are in Hardy–Weinberg equilibrium; 88% of the 1008 heterozygosities (63 markers in 16 populations) are greater than 9.5%. As expected, average heterozygosities vary both for the different loci and the different populations (Table 1). For all but one of the markers (#12 at *TNFRSF6*) the F_{ST} values fall within the bulk of the distribution shown by more than 200 SNPs studied on these populations (Table 1) (see Supplementary Material Figures S5, S6).

Linkage disequilibrium

Figure 1 shows the quantitative patterns of LD for the 16 populations across each of the three loci. For some regions, the patterns of LD along these loci show substantial variation among the populations. Given the sample sizes and heterozygosities, most ξ values of 0.3 or greater are statistically significantly different from 0.0 at $P<0.001$. Thus, all three loci have segments across which LD is strong and highly significant in all populations, as well as segments across which no population shows significant LD. However, for most of the segments, there is considerable variation among populations in the magnitude of LD and the relative rankings are not consistent, although there is a trend for African populations to have the least LD and the 'Eastern' (east Asian, Pacific, and Native American) populations to have the greatest LD.

All three loci have regions in which the magnitude and pattern of LD vary considerably among the populations. Populations within a geographic region tend to be similar, but differences occur among the geographic regions. At *RET-D10S94*, a region of at least 80 kb (from ~40 to 120 kb) shows high LD for all non-African populations but practically no LD for any African population, except the Ethiopians, who showed intermediate levels of LD for part of the region. Across two adjacent intervals at *TNFRSF6* (from ~40 to 80 kb) all populations show elevated LD but the range is very large from $\xi = 0.3$ to 2.4. At *DRD2-NCAM1*, there is one region of ~70 kb (from ~80 to 150 kb) where high LD is shown only by the Japanese and the other 'Eastern' populations, while all African and European populations show low LD across this region.

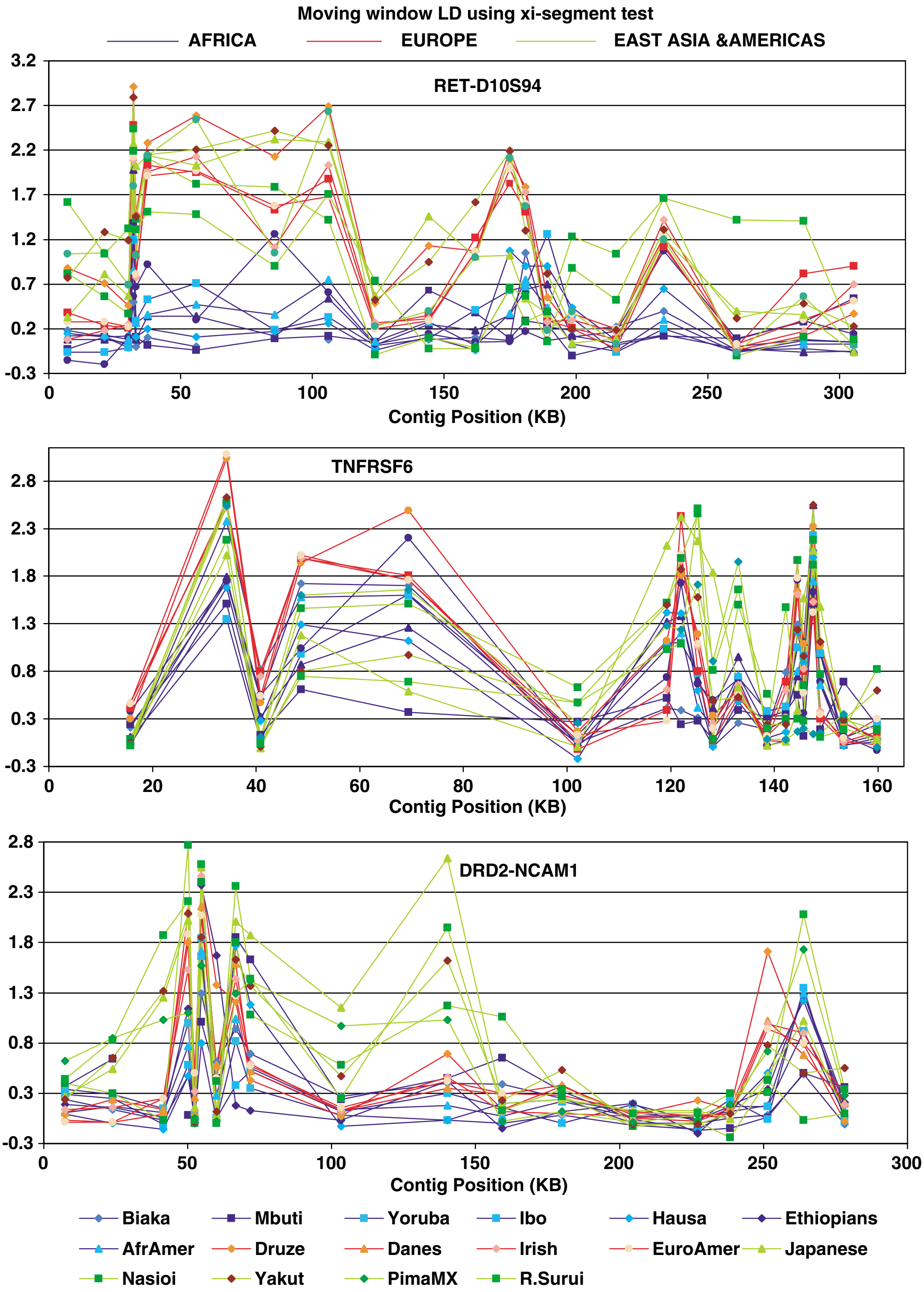

Figure 1 The quantitative patterns of LD along each of the three regions. Values of ξ, calculated as a segment test, are plotted at the midpoints of the segments. The lines connecting points for the same population are colored by geographic region: shades of blue for African populations, red to orange for European populations, and shades of green for 'Eastern' populations (East Asia, Pacific, and Native American).

We also see some regions where the pattern of LD differs among populations within a geographic region. There is a segment of ~75 kb at *RET-D10S94* (from ~125 to 200 kb) where differences in the pattern of LD are seen among populations from the same geographical region. At one small segment at *TNFRSF6* (from 120 to 124 kb), there is significant heterogeneity among the African populations and between the American and East Asian populations. At *DRD2-NCAM1*, a small region of 30 kb (from ~245 to 275 kb) shows differences in patterns of LD (see Supplementary Material). As no uniformity is seen for LD in populations even from the same geographical region, an LD map for such genomic regions cannot be generalized.

Haplotypes

The LD profiles in Figure 1 do not allow inference of the haplotypes or their frequencies. In contrast, the LD profiles can be generated from the haplotypes and their frequencies. Thus, the underlying estimated haplotypes and their frequencies are the primary data; LD is a statistical abstraction. For each of the three genetic regions studied, most of the haplotype frequency distributions were very significantly different from one another ($Pr < 0.001$) when populations were compared pairwise. For example, in the *TNFRSF6* region almost 80% of the unique pairwise population comparisons were significantly different ($Pr < 0.010$) in the series of four-site moving window haplotypes generated and 70.5% of the comparisons had probabilities < 0.001. Typically, the haplotype frequency distributions that were not statistically different from one another were for population sample comparisons from within the same geographical region, but even then many of the within region comparisons were also significantly different.

The relationships between LD and haplotype frequencies are complex; the following examples serve to illustrate this complexity. Haplotype frequencies for the common haplotypes in the 16 populations for selected segments of each genomic region are given in Figure 2 adjacent to the LD profile for that segment. At *RET-D10S94*, there are five common haplotypes for five SNPs, numbers 5–9 (mapping from 33 to 105 kb in Figure 1). The LD is reasonably uniform across the four intervals with significant quantitative variation among the populations with African populations low and European and 'Eastern' populations high (Figure 2a). The same few haplotypes are the common ones in almost all populations, but the frequencies differ considerably, even among the non-African populations (Figure 2b). The marked difference in LD between African and non-African populations is not a difference in what haplotypes are present but a difference in the frequencies of those few haplotypes. At *TNFRSF6*, seven haplotypes, of which only three are globally common, defined by SNPs numbers 2–5 (mapping from 30 to 60 kb in Figure 1) account for almost all chromosomes in all populations. The expanded version of this region shows a nearly identical pattern of LD for all populations, but considerable quantitative variation among the populations across the region as well as variation across the different subsegments of the region (Figure 2c). That quantitative variation, however, is not immediately obvious from the haplotype frequencies in Figure 2d. In the African and European populations, the same three haplotypes predominate with relatively minor variation in haplotype frequencies, whereas the frequencies are considerably different for the 'Eastern' populations. At *DRD2-NCAM1*, the region from TaqI 'D' to SNP920 extends from 56 to 132 in Figure 1. The expanded version of this region shows considerable quantitative variation in nonrandomness among populations, but very similar patterns across the region (Figure 2e). The haplotypes of these five SNPs, however, show large differences among populations (Figure 2f). Although European and African populations have very different haplotype frequencies, LD tends to be low for both groups. The most common African haplotype is uncommon elsewhere and the most common European haplotype is rare, or absent, in most other populations. The 'Eastern' populations have yet a different set of most common haplotypes and show generally high LD.

These differences point out the distinction, often overlooked, between measures of LD and the underlying estimated haplotype frequencies: differences in the amount of LD can be generated by different frequencies of the same few haplotypes, and similar amounts of LD can be generated by very different sets of haplotype frequencies. Additional examples are given in the Supplementary Material.

Discussion

Most of the population-specific allele frequencies were higher than 5% (Table 1). Since worldwide populations were studied, it is quite reasonable that allele frequencies will be very different among populations due to the effect of random genetic drift. Markers with low heterozygosities are not useful for estimating LD and, therefore, ξ values were estimated across the segments using flanking pairs of markers so as to compensate for the effect of the occasional low heterozygosity at a single marker. As noted earlier, this approach also incorporates nonrandomness shown by nearby nonadjacent markers.

LD is inherently a population genetics and statistical measure. The underlying biological data are the haplotype frequencies in the populations. The EM algorithm gives accurate estimates of the frequencies of common haplotypes,[47–49] especially when there is significant disequilibrium. These haplotype frequencies can also provide information on evolutionary histories, beyond what can be learned from individual markers.[5,7] Fundamental bio-

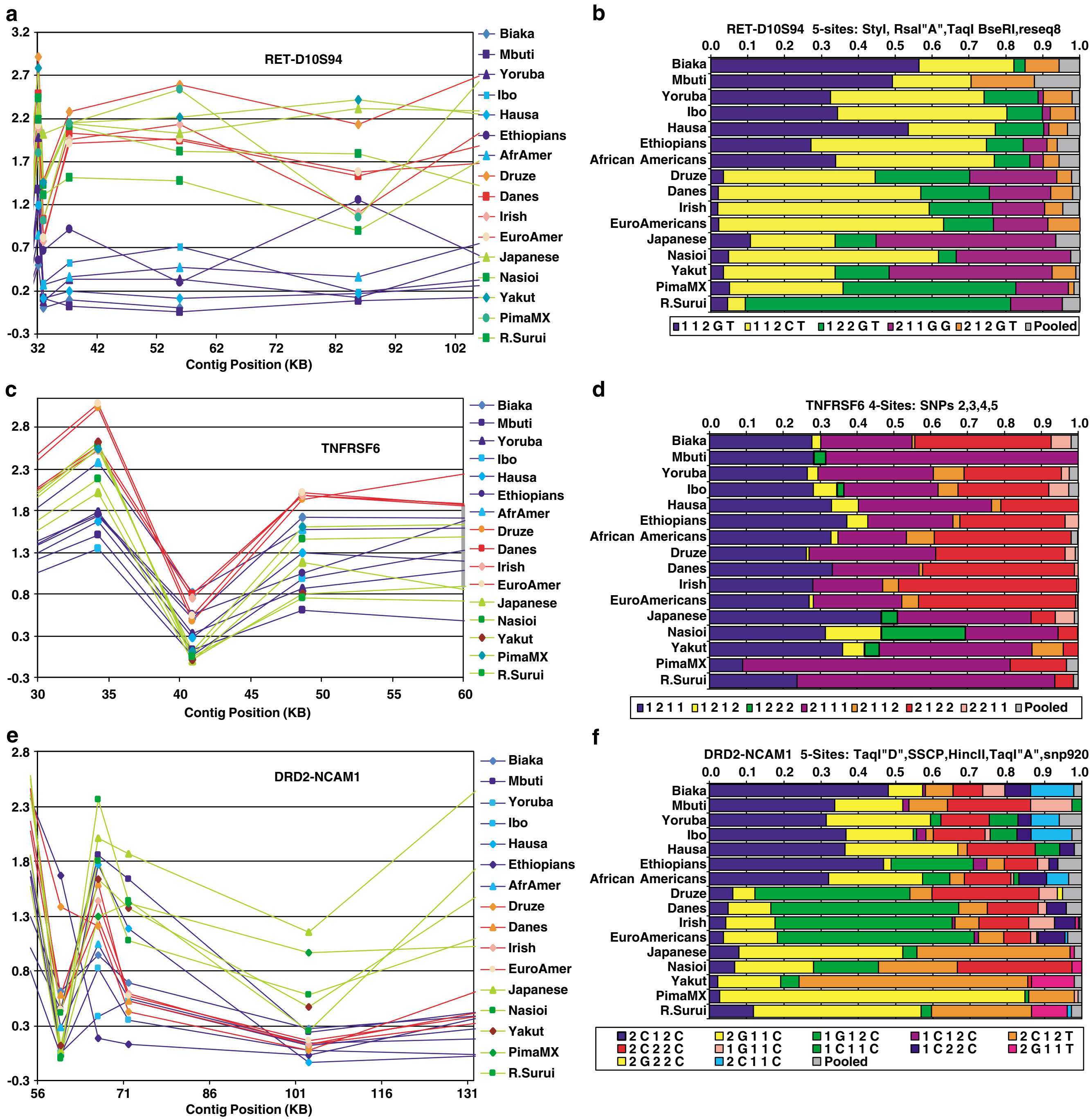

Figure 2 Comparisons of LD patterns for subsegments of each locus with the haplotypes and their frequencies for the same subsegment. For each of the LD plots, the contig position is the same as in Figure 1 for the corresponding locus and the graphic conventions are as described for Figure 1. The haplotype frequencies are represented as stacked bars with length corresponding to frequency and the haplotypes colored arbitrarily. The gray 'pooled' category includes all haplotypes never seen at frequencies greater than 5% in any of the populations. This category includes the haplotypes inferred to be present at frequencies of less than one chromosome in the sample because of occasional missing data for one of the sites. (**a** and **b**) LD and haplotype frequencies for a segment of the *RET-D10S94* region. (**c** and **d**) LD and haplotype frequencies for a segment of the *TNFRSF6* region. (**e** and **f**) LD and haplotype frequencies for a segment of the *DRD2-NCAM1* region.

logical processes such as mutation and recombination[50] are important but not the sole factors determining LD. Population demographic history, through its impact on random genetic drift, may be the major factor for determining most of the LD patterns in humans. Especially relevant will be the interaction between rates of recombination and rates of random genetic drift. The statistic used to measure LD is an important aspect of the problem because different statistics assess nonrandomness of alleles, based on the haplotype frequencies, in different ways, and consequently yields different views of the underlying population genetics.

Most commonly used statistics assess LD in a pairwise manner between two diallelic polymorphisms. D', commonly used to measure LD, has absolute values ranging from zero to 1. $|D'|=1$ indicates the absence of any evidence of historical recombination between the two sites. The four-gamete test is equivalent to testing whether $|D'|<1$ and thereby indicates that recombination has occurred between the two sites, and that haplotypes (at least 1) that are descended from a meiotic crossover have been observed (note, however, that the inference of a crossover assumes recombination is a more likely explanation than recurrent mutation). Also, it is not possible to infer from $|D'|<1$ how frequently recombination occurred or when in the past a crossover occurred. Consider that all polymorphisms have frequent alleles because a mutation occurred and descendant copies of the new form became common due to random genetic drift or selection operating on the new variant or on a nearby site (hitchhiking). Similarly, a very uncommon crossover event, giving rise to one copy of the 'new' haplotype, could have descendant copies at high frequency. Of course, the more likely the crossovers, the more likely some of the crossover descendant haplotypes will be observed in a sample of chromosomes in a population. However, there is a large chance component such that in a specific case one cannot equate haplotype frequencies with recombination rate between the sites. Other measures of LD are less closely related to the question of 'obligate' crossovers and are more closely correlated with whether the haplotype frequencies differ from what would be expected by chance.[44] How best to assess LD and which measure is best for which research question are areas of active interest.

Other measures have also been applied to the data. Starting from either end, we can measure the 'decline' of LD as more distant markers are considered in pairwise tests. Any of several of the pairwise statistics can be used; D' and Δ^2 have been considered here. We have considered these two measures in two directions, and anchoring on either the last or the next to the last marker for each direction (data not shown). The conclusions are several. The decline pattern is very different for the two statistics and can be very different across the same regions depending on which marker is the anchor. Also worth noting is that different populations, even those from the same geographic region, can show similar patterns in some genomic regions and very different patterns in other regions. We do not think averages across these very different patterns are meaningful, but we do believe that some generalizations are possible. We consistently see low LD among the African populations at all three loci in this study. This generalizes the previous findings at several other loci on a few African populations[12,23] and a few specific smaller loci on several African and non-African populations.[5,8] The low LD among the African populations can be clearly attributed to the larger long-term effective population size for African populations because of the African origin of modern humans. On the other hand, high LD shown by the non-African populations is explainable by the relatively short time span subsequent to the founder event associated with the expansion out of Africa such that relatively little recombination has occurred. At the *DRD2-NCAM1* locus, we find intervals where there is no LD for any of the populations. We searched, unsuccessfully, for the presence of any consensus sequence for recombination or difference in genetic composition of these intervals, and so we conclude there are other factors operating in these intervals.

The three genomic regions have different relative chromosomal locations – the *RET-D10S94* is centromeric while *TNFRSF6* and *DRD2-NCAM1* are in the middle of the long arms of chromosome 10 and 11, respectively – and thus are expected to show different overall levels of LD because of the lower frequency of recombination in general in centromeric regions. The centromeric locus does show one region of high LD in non-Africans (from ~40 to 110 kb) that is longer than any in the other two loci, which is as expected, but there are also regions of low LD, just as in the other loci. There are reports saying that LD may extend up to 500 kb.[26,27] We cannot evaluate such long-range LD in the existing data but do note weak but significant pairwise LD across the extremes in five instances (Supplementary Material Table S2). Three instances occurred in the Rondonian Surui: between the penultimate markers at *TNFRSF6*, and between the first two markers and the penultimate marker in the *RET-D10S94* region. Significant LD also occurred in the Biaka: between the first and last markers at *DRD2-NCAM1*, and in the Druze: between the second and the last SNPs at *TNFRSF6*. Thus, long-range LD in these regions is absent in most populations and occurs idiosyncratically in the more isolated populations.

There are some regions where the LD pattern is very different in different populations. Also, the populations belonging to a particular geographic region do not show consistency in pattern of LD for all the three loci. We have observed that the extent of disequilibrium shown by African and European populations are similar at the *DRD2-NCAM1* locus and shorter than seen for the East Asian, Pacific islanders, and New World populations. At the *RET-D10S94* locus, the LD value is always very low for the Africans, intermediate for the Europeans, and very high among the East Asian, Pacific Islanders, and American populations. At the *TNFRSF6* locus, the Europeans in general exhibit a higher level of disequilibrium than the other non-African populations, while the African populations are at their usual low. Most of the SNPs typed are in noncoding regions, and all but one of those in the coding regions are synonymous changes and therefore likely free from the direct effects of selection. Three of the synonymous SNPs and the one nonsynonymous SNP have F_{ST} values well within the general F_{ST} distribution (Table 1 and Supplementary Material). So, we conclude that variation observed at all but one site (*TNFRSF6* site #12) is due to population history.

TNFRSF6 site 12 (rs2229521) is the only one of the 63 markers with an unusual F_{ST} value. That value is entirely attributable to high frequencies (0.57 and 0.68) in the two Native American populations for an allele that is absent in all native Africans and present in other populations at no more than 4%. This could just be an extreme example of random genetic drift associated with the founding of Native Americans. However, the question arises whether this may be due to selection on this SNP (a synonymous coding SNP) or an untested marker in strong LD with it. Such selection could be restricted to the Americas or involve an allele that arose in the population ancestral to the Native Americans. The flanking markers do not show such elevated F_{ST} values but do show significant LD with marker 12 in these two Native American populations (see Supplementary Material). The data are compatible both with selection having increased the frequency of this haplotype in (the founders of) Native Americans AND with the generally greater extent of LD in Native American populations because of a bottleneck associated with the initial colonization.[30]

Our analysis of 16 globally representative populations has demonstrated that allele frequencies and common haplotypes differ between populations, even between those from similar geographical origins. The result is considerable qualitative and quantitative variation in patterns of LD among populations. Considering that global populations have different demographic histories, and their genomes have been shaped differently by factors such as drift, recombination, and mutation, this diversity is not surprising. We do not see strong evidence of block-like structures across the regions studied, but the ξ statistic is not designed to identify blocks. We do see regions with more as well as less LD and these regions could by some definitions be called blocks. Application of a block-finding algorithm to the *TNFRSF6* data yielded a very complex pattern of similarities and differences among populations.[51] The same analysis also showed that for these data tagging SNPs differed among the populations, even among populations from the same geographic region.[51] Simple inspection of some of the haplotype data at *DRD2-NCAM1* (Figure 2f), for example, shows that tagging SNPs need to distinguish different haplotypes in different regions of the world. As of redundancy between some SNPs, it is possible in this case to select a set of tagging SNPs for an African population that will or will not be appropriate for populations in other regions. Thus, in some cases the same SNPs will work to distinguish very different sets of haplotypes. In other cases, no common minimal set of SNPs will suffice to distinguish among different sets of haplotypes. The point is that it is impossible to know *a priori*; it will depend on each specific case.

Our objective is to emphasize the variation seen from some regions of high LD in all populations through regions of large variation in LD among populations to regions of low LD in all populations. This variation will translate to different sets of tagging SNPs for different populations.[51] While the vast amount of LD and haplotype data gathered by the HapMap project will certainly be a useful starting point, it will be important to assess the utility of the general map(s) to the specific population of interest before embarking on disease genetic studies.

Acknowledgements

We would like to thank William C Speed for his technical help. This work was supported in part by US PHS grants, MH62495 and GM57672, to KKK, the Swedish Society for Medical Research, and the Center for Genomics and Bioinformatics departmental funds provided by Pfizer (formerly Pharmacia). We would like to acknowledge and thank the following additional individuals for their help over the years in assembling the samples from the diverse populations studied: FL Black, B Bonne-Tamir, LL Cavalli-Sforza, J Friedlaender, E Grigorenko, K Kendler, A Odunsi, F Okonofua, F Oronsaye, J Parnas, LO Schulz, and OV Zhukova. Additional samples were obtained from the National Laboratory for the Genetics of Israeli Populations and from the Coriell Institute for Medical Research. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies such as this. Without such participation of individuals from diverse parts of the world, we would be unable to obtain a true picture of the genetic variation in our species.

References

1 Johnson GC, Esposito L, Barratt BJ *et al*: Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29**: 233–237.
2 Kruglyak L: Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 1999; **22**: 139–144.
3 Weiss KM, Clark AG: Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002; **18**: 19–24.
4 Kidd KK, Morar B, Castiglione CM *et al*: A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum Genet* 1998; **103**: 211–227.
5 Tishkoff SA, Dietzsch E, Speed W *et al*: Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 1996; **271**: 1380–1387.
6 Tishkoff SA, Goldman A, Calafell F *et al*: 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* 1998; **62**: 1389–1402.
7 Tishkoff SA, Pakstis AJ, Stoneking M *et al*: Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: implications for modern human origins. *Am J Hum Genet* 2000; **67**: 901–925.
8 Kidd JR, Pakstis AJ, Zhao H *et al*: Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in

a global representation of populations. *Am J Hum Genet* 2000; **66**: 1882–1899.

9 Oota H, Pakstis AJ, Bonne-Tamir B *et al*: The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. *Ann Hum Genet* 2004; **68**: 93–109.

10 Collins FS, Brooks LD, Chakravarti A: A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* 1998; **8**: 1229–1231.

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

12 Reich DE, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.

13 Cardon LR, Abecasis GR: Using haplotype blocks to map human complex trait loci. *Trends Genet* 2003; **19**: 135–140.

14 Goldstein DB: Islands of linkage disequilibrium. *Nat Genet* 2001; **29**: 109–111.

15 Kauppi L, Sajantila A, Jeffreys AJ: Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum Mol Genet* 2003; **12**: 33–40.

16 Jeffreys AJ, Kauppi L, Neumann R: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 2001; **29**: 217–222.

17 Jeffreys AJ, Holloway JK, Kauppi L *et al*: Meiotic recombination hot spots and human DNA diversity. *Philos Trans R Soc London Ser B* 2004; **359**: 141–152.

18 Gibbs RA, Belmont JW, Hardenbol P *et al*: The International HapMap Project. *Nature* 2003; **426**: 789–796.

19 Couzin J: Genomics. New mapping project splits the community. *Science* 2002; **296**: 1391–1393.

20 Phillips MS, Lawrence R, Sachidanandam R *et al*: Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet* 2003; **33**: 382–387.

21 Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L: Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. *Hum Genet* 2003; **113**: 51–59.

22 Patil N, Berno AJ, Hinds DA *et al*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294**: 1719–1723.

23 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.

24 Taillon-Miller P, Bauer-Sardina I, Saccone NL *et al*: Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat Genet* 2000; **25**: 324–328.

25 Dawson E, Abecasis GR, Bumpstead S *et al*: A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 2002; **418**: 544–548.

26 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.

27 Abecasis GR, Noguchi E, Heinzmann A *et al*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001; **68**: 191–197.

28 Tsunoda T, Lathrop GM, Sekine A *et al*: Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. *Hum Mol Genet* 2004; **13**: 1623–1632.

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

30 Kidd KK, Pakstis AJ, Speed WC, Kidd JR: Understanding human DNA sequence variation. *J Hered* 2004; **95**: 406–420.

31 Jorde LB, Watkins WS, Kere J, Nyman D, Eriksson AW: Gene mapping in isolated populations: new roles for old friends? *Hum Hered* 2000; **50**: 57–65.

32 DeMille MM, Kidd JR, Ruggeri V *et al*: Population variation in linkage disequilibrium across the COMT gene considering promoter region and coding region variation. *Hum Genet* 2002; **111**: 521–537.

33 Kamatani N, Sekine A, Kigamoto T *et al*: Large scale single nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP Maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. *Am J Hum Genet* 2004; **75**: 190–203.

34 Zhao H, Pakstis AJ, Kidd JR, Kidd KK: Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. *Ann Hum Genet* 1999; **63**: 167–179.

35 Nothnagel M, Furst R, Rohde K: Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered* 2002; **54**: 186–198.

36 Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK: ALFRED: an allele frequency database for anthropology. *Am J Phys Anthropol* 2002; **119**: 77–83.

37 Chen TJ, Boles RG, Wong LJ: Detection of mitochondrial DNA mutations by temporal temperature gradient gel electrophoresis. *Clin Chem* 1999; **45**: 1162–1167.

38 Howell WM, Jobs M, Gyllensten U, Brookes AJ: Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. *Nat Biotechnol* 1999; **17**: 87–88.

39 Livak KJ: Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. *Genet Anal* 1999; **14**: 143–149.

40 Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ: HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res* 2002; **30**: 387–391.

41 Wright S: Evolution and the genetics of populations; in *The Theory of Gene Frequencies*. Chicago, IL: University of Chicago Press, Vol. 2,: 1969, p 511.

42 Hawley ME, Kidd KK: HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* 1995; **86**: 409–411.

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

44 Devlin B, Risch N: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29**: 311–322.

45 Maniatis N, Collins A, Xu CF *et al*: The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci USA* 2002; **99**: 2228–2233.

46 Workman PL, Niswander JD: Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. *Am J Hum Gen* 1970; **22**: 24–49.

47 Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK: The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. *Am J Hum Genet* 2000; **67**: 518–522.

48 Zhang S, Pakstis AJ, Kidd KK, Zhao H: Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *Am J Hum Genet* 2001; **69**: 906–914.

49 Fallin D, Schork NJ: Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 2000; **67**: 947–959.

50 Greenwood TA, Rana BK, Schork NJ: Human haplotype block sizes are negatively correlated with recombination rates. *Genome Res* 2004; **14**: 1358–1361.

51 Liu N, Sawyer SL, Mukherjee N *et al*: Measures of similarity for haplotype block structures show variation among populations. *Genet Epidemiol* 2004; **27**: 385–400.

Supplementary Information accompanies the paper on the European Journal of Human Genetics website (http://www.nature.com/ejhg)