Linkage disequilibrium in human populations

Christine Lonjou^{*†}, Weihua Zhang^{**}, Andrew Collins^{*}, William J. Tapper^{*}, Eiram Elahi^{*}, Nikolas Maniatis^{*}, and Newton E. Morton^{*§}

^{*}Human Genetics Division, University of Southampton, Duthie Building (MP 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, United Kingdom; and [†]Institut National de la Santé et de la Recherche Médicale U563, Avenue de Grande Bretagne, 31059 Toulouse, France

Contributed by Newton E. Morton, March 10, 2003

Whereas the human linkage map appears on limited evidence to be constant over populations, maps of linkage disequilibrium (LD) vary among populations that differ in gene history. The greatest difference is between populations of sub-Saharan origin and populations remotely derived from Africa after a major bottleneck that reduced their heterozygosity and altered their Malecot parameters, increasing the intercept M that reflects association in founders and decreasing the exponential decline ϵ . Variation among populations within this ethnic dichotomy is much smaller. These observations validate use of a cosmopolitan LD map based on a sizeable sample representing a large population reliably typed for markers at high density. Then an LD map for a region or isolate within an ethnic group may be created by fitting the sample LD to the cosmopolitan map, estimating Malecot parameters simultaneously. The cosmopolitan map scaled by ϵ recovers 95% of the information that a local map at the same density gives and therefore more than the information in a low-resolution local map. Relative to a Eurasian cosmopolitan map the scaling factors are estimated to be 0.82 for isolates of European descent, 1.53 for Yorubans, and 1.74 for African Americans. These observations are consistent with a common bottleneck (perhaps but not necessarily speciation) $\approx 173,500$ years ago, if the bottleneck associated with migration out of Africa was 100,000 years ago. Eurasian populations (especially isolates with numerous cases) are efficient for genome scans, and populations of recent African origin (such as African Americans) are efficient for identification of causal polymorphisms within a candidate sequence.

The density of known polymorphisms in the DNA sequence is the pacemaker for studies of population structure. Blood groups and later isozymes provided enough information to bioassay kinship at single loci, which was confirmed by studies of migration, genealogy, and isonymy (1, 2). Then DNA markers at increasing density permitted analysis of allelic association between pairs of loci, also called linkage disequilibrium (LD). An important application replaces one locus by a phenotype of interest, with the intent of localizing a gene of unknown sequence contributing to that phenotype. This enterprise is called positional cloning. A beginning has been made on haplotype diversity, treating haplotypes as alleles and measuring their diversity by heterozygosity or Shannon's entropy (3). With all approaches the arbitrary choice of an arbitrary number and type of markers has discouraged use of haplotypes to describe population structure, although they are of value for positional cloning of major genes with short history and nearly monophyletic origin (4). It remains to be seen how haplotype annotation of LD will facilitate positional cloning, especially of oligogenes contributing to multifactorial diseases.

Whether markers are studied individually or in haplotypes, population structure can aid positional cloning by providing replicates for LD, perhaps using populations with high LD to identify a candidate region and then populations with less LD to narrow the support interval. To evaluate these possibilities we must establish the degree and pattern of LD in different populations, which we investigate here.

Materials and Methods

The data consist of studies on diallelic polymorphisms that compared two or more populations. Most data are for short

sequences at low resolution, and the populations are not well characterized. Does a sample represent an extended family or a region? How could a region be randomly sampled? If close relatives were avoided, how was selection carried out, given that an isolate may have more than a dozen chains of relationship between two random members? Cell and DNA banks do not address such questions. Studies of LD are in their infancy, and population comparisons lag behind single samples of ill-defined origin.

Some insight is provided by theory. The variance from additive genetic effects is $V_G(1 - \alpha)$ in two random populations, $V_G(1 + \alpha)$ in their F_1 , and V_G in the F_2 . Neglecting genetic drift and introgressive hybridization, the expectation for F_n remains V_G . Inbreeding α is <0.01 in all large populations that do not prefer consanguineous marriage (5) and rarely reaches 0.1 even in highly inbred isolates (2). Among populations large enough to be useful for positional cloning of oligogenes, we do not expect α to be consequential. Of course, the situation is different for rare alleles. If q is the gene frequency in a large ancestral population, the probability that the allele in question was absent from an isolate that originated with N founders is e^{-2Nq} , or $\approx 1 - 2Nq$ if Nq is small. In the unlikely event that the allele was present, the initial frequency was at least $1/2N$ and might through drift become greater. This makes isolates a rich reservoir of rare alleles inherited from a single founder in the not-far-distant past, which are not major contributors to common diseases.

The theory for pairwise LD is more complicated than theory for single loci, and extension to multilocus haplotypes has not yet been successful. At small distances a close approach to equilibrium requires constant effective size over thousands and perhaps millions of years as LD from founders is diminished by recombination but increased by drift, making haplotype frequencies inconstant in time. Allele frequencies are also inconstant, but the possibility that a current polymorphism might have been monomorphic is not a useful guide to LD (6). Attempts to predict the squared value of LD as its ratio to an expected function of allele frequencies have also been unsuccessful, because their metric is not a probability, does not predict haplotype frequencies, and has no evolutionary theory (7). For two diallelic loci the optimal measure of allelic association is the probability ρ , which estimates haplotype frequencies, has an evolutionary theory and minimal variance from expectation, and is applicable to both case-control and random samples (8). The expected value of ρ follows the Malecot equation $(1 - L)Me^{-\epsilon d} + L$, where the asymptote L measures association at large distance and the intercept M reflects association in founders.

The ρ metric has a bias L with predicted value L_p equal to the K_p -weighted mean of $\sqrt{2/\pi K_p}$, where K_p , the information about ρ per marker pair, is proportional to sample size (9). L_p does not depend on the Malecot model, but only on the mean value of ρ for markers at distances so great that the expected value of disequilibrium D is zero, under the assumption that the mean

Abbreviations: LD, linkage disequilibrium; LDU, LD unit; SNP, single-nucleotide polymorphism.

^{*}C.L. and W.Z. contributed equally to this work.

[§]To whom correspondence should be addressed. E-mail: nem@soton.ac.uk.

Table 1. Means over the MNS, RHCE, RHCD, RHDE, and CD4 allotypes

	ρ	Het	Link	Yule		ρ	Het	Link	Yule
Regions (39 pooled samples)					F ₁ with European (34 pooled samples)				
Europe	0.780	0.652	0.230	0.812					
Near East	0.778	0.656	0.226	0.799	Near East	0.789	0.649	0.204	0.821
India and Pakistan	0.698	0.560	0.160	0.718	India and Pakistan	0.750	0.612	0.184	0.782
East Asia	0.526	0.471	0.111	0.564	East Asia	0.738	0.597	0.175	0.769
Sub-Saharan Africa	0.477	0.512	0.064	0.484	Sub-Saharan Africa	0.713	0.623	0.126	0.744
Amerindians	0.720	0.541	0.152	0.744	Amerindians	0.760	0.624	0.192	0.796
Oceania	0.643	0.294	0.040	0.660	Oceania	0.753	0.555	0.155	0.817
North Africa	0.773	0.643	0.191	0.820	North Africa	0.781	0.647	0.196	0.830
Mean	0.674	0.541	0.146	0.700	Mean	0.755	0.615	0.176	0.794
SE	0.043	0.022	0.018	0.043	SE	0.036	0.010	0.013	0.032
Isolates (24 pooled samples)					Ratios				
Basques	0.792	0.606	0.196	0.836	Isolates/region	1.18	1.12	1.32	1.17
Jews	0.780	0.624	0.176	0.799	Europe/isolate	0.98	1.07	1.20	0.97
Eskimos	0.831	0.579	0.166	0.833	F ₁ /region	1.12	1.14	1.21	1.13
Lapps	0.800	0.597	0.184	0.829	Europe/F ₁	1.03	1.06	1.13	1.02
Ainus	0.846	0.596	0.260	0.898					
Tristan Da Cunha	0.711	0.641	0.170	0.728					
Mean	0.793	0.607	0.192	0.821					
SE	0.057	0.012	0.022	0.053					

Het, heterozygosity; Link, linkage information; ρ and Yule, LD metrics.

value of K_ρ is independent of distance and that ρ for given K_ρ is distributed as the absolute value of a normal deviate with variance $1/K_\rho$. On the contrary, the simple Malecot model with constant ε assumes parsimony that is violated by blocks and steps, giving a better representation of LD in a long sequence than in a short one. The likelihood under any model cannot provide an optimal test of a subhypothesis when the general hypothesis is wrong. We have found that a local effect of block structure can distort the estimate of L in a short sequence, inflating steps in the LD map. Therefore $L = L_p$ was assumed even when an estimate of L made a significant improvement in likelihood. Distance d is additive and may be measured in kb if association is homogeneous on the physical map, in centimorgans if association is determined by recombination and the genetic map is reliable, or in LD units (LDU), where $\varepsilon = 1$ when d is replaced by $\sum \varepsilon_i d_i$, the inner product of distance in kb between i th adjacent markers and the corresponding local value of ε . Then 1 LDU corresponds to $\sum \varepsilon_i d_i = 1$, and the spanned distance in kb is a swept radius within which LD declines to e^{-1} of the intercept (9).

There is a substantial gain in efficiency with the LD map, which allows for hot and cold spots that reflect recombination,

a selective sweep, or chance (10–12). To construct LD maps we used the interval method (9) as modified (12). The map for a given population may be based entirely on samples from that population or on a cosmopolitan standard that concatenates samples of single-nucleotide polymorphism (SNP) pairs from two or more populations, varying among studies. For each genomic region within a study we combined estimates of ρ over samples weighted by information on the null hypothesis that $\rho = 0$ (9). This cosmopolitan map in LDUs with $\varepsilon = 1$, M , L_p was used to estimate the Malecot parameters for each sample. The error variance V for m pairs of SNPs with composite likelihood $\text{lk} = \exp[-\sum K_\rho(\hat{\rho} - \rho)^2/2]$ was taken as $V = -2\ln \text{lk}/\text{df}$, where V and the degrees of freedom (df) depend on the model. The nominal variance of any estimate $\hat{\theta}$ is $V\sigma_\theta^2$, where σ_θ^2 was taken from the covariance matrix calculated from second derivatives that do not include V . The corresponding information K_θ is the reciprocal of $V\sigma_\theta^2$, and the standard error is $\sqrt{V\sigma_\theta^2}$. Composite likelihood gives consistent estimates and good search intervals, but the standard errors and confidence intervals are less reliable (13). To pool s samples we took their mean $\sum \theta_j/s$ with information $s^2/\sum(1/K_{\theta_j})$. To compare two samples or groups of samples the difference $\theta_j - \theta_k$ was assigned information $1/(1/K_{\theta_j} + 1/K_{\theta_k})$. To make the

Table 2. Samples analyzed

Population	Origin	Isolate	Eaves <i>et al.</i> (15)	Taillon-Miller <i>et al.</i> (16)	Dunning <i>et al.</i> (17)	Gabriel <i>et al.</i> (18)
Yoruban	Africa	0	0	0	0	+
African-American	Africa	0	0	0	0	+
Japanese/Chinese	Asia	0	0	0	0	+
European	Europe	0	0	0	0	+
CEPH	Europe	0	0	+	0	0
American	Europe	0	+	0	0	0
British	Europe	0	+	0	+	0
Finnish	Europe	+	+	+	+	0
Sardinian	Europe	+	+	+	0	0
Afrikaner	Europe	*	0	0	+	0
Ashkenazim	Europe	+	0	0	+	0

CEPH, Europeans from the Centre d'Etude du Polymorphisme Humain.

*, Analysed as isolate, but status of this sample is equivocal (Table 7).

Table 3. Degrees of freedom for alternative maps

Parameters	Concatenated samples j	Total	Individual samples i	Allocated to i th sample within j
No. of pairs	N_j		N_i	
No. of markers	m_j		m_i	
No. of samples	J		I	
Type of map				
Cosmopolitan, kb		$\sum_j(N_j - 3)$		$N_j - 3 (i = j)$
Cosmopolitan, LDU		$\sum_j[N_j - (m_j + 2)]$		$N_j - (m_j + 2) (i = j)$
Samples with cosmopolitan LD map ($\hat{M}, \hat{\epsilon}$)		$\sum_{i,j}[N_i - (m_j - 1) - 3/I]$		$N_i - (N_i/N_j)(m_j - 1) - 3$
As above, M from cosmopolitan map		$\sum_{i,j}[N_i - (m_j - 1) - 2/I]$		$N_i - (N_i/N_j)(m_j - 1) - 2$
Samples with own map		$\sum_i[N_i - (m_i + 2)]$		$N_i - (m_i + 2)$

When samples contributing to a cosmopolitan map differed in numbers of markers, $m_j - 1$ and $m_j + 2$ were distributed among samples in proportion to N_j for the purpose of calculating a sample-specific variance.

exponent of M comparable to ϵ , we took $\theta_j = \epsilon_j$, $-\ln M_j$ and $\sigma_{-\ln M_j}^2 = \sigma_{M_j}^2/M^2$, with M from the cosmopolitan map. Then the expected values of both parameters are related to time since the founder population (8). We also used these methods to estimate ϵ_j conditional on the value of M in the cosmopolitan map to test whether this simpler analysis is equally sensitive to population differences and justifies scaling of local maps by ϵ_j alone. Relative efficiency of alternative maps was defined as the ratio of their variances, which for multiple samples were estimated by summing their quadratic forms and degrees of freedom. Genetic analysis was performed with ALLASS and LDMAP programs (<http://cedar.genetics.soton.ac.uk/public.html/>).

Single Pairs of Markers in Regions, Isolates, and F₁ Hybrids

Lonjou *et al.* (14) compared 1,266 samples from published databases. The data consisted of haplotype frequencies estimated for pairs of diallelic markers in five loci. Within a pair, the rarest of the four haplotypes was not consistent among populations. This finding alters the definition of ρ , but the absolute value of the correlation is unaffected and was therefore preferred, although subsequently shown to be inefficient (8). Here we impose consistency by pooling samples within eight regions and six isolates, from which expected haplotype frequencies for an F₁ between Europeans and other regions are calculated, together with the efficient ρ metric and three other statistics from the haplotype frequencies ($f_{11}, f_{12}, f_{21}, f_{22}$) arranged conventionally so that $D = f_{11}f_{22} - f_{12}f_{21} \geq 0$ and $f_{21} \geq f_{12}$. Letting $f_{11} + f_{21} = R$, $f_{11} + f_{12} = Q$, and therefore $Q \leq R$, $Q(1 - R) \leq QR$, $R(1 - Q)$, $(1 - Q)(1 - R)$: $\rho = D/Q(1 - R)$, $\text{het} = 1 - \sum_{ij} f_{ij}^2$, $\text{link} = 2(f_{11}f_{22} + f_{12}f_{21})$, and $\text{yule} = D/\text{link}$. The CD4 locus was not typed in all regions and isolates. Where present it was adjusted by linear regression so that its mean equaled the mean of the other loci.

Table 1 shows that heterozygosity is significantly higher in the specified isolates than in regions, contrary to expectation, but the differential is only 12%. Linkage information shows a larger but nonsignificant excess in isolates. The two LD metrics (ρ and yule)

have a similar (nonsignificant) increase in isolates. Sub-Saharan Africa shows markedly reduced LD compared with the other regions, consistent with a population bottleneck during or subsequent to migration out of Africa. In contrast with isolates, the haplotype frequencies for an F₁ between Europeans and people from other regions give means close to Europeans. Whereas isolates probably offer some advantage for detection of LD, interracial crosses are less predictable for both linkage and LD. This preliminary analysis confounds the Malecot parameters and therefore cannot estimate scaling factors for an LD map, but it encourages examination of more informative data on syntenic markers.

Multiple Markers in Regions and Isolates

Several studies provide closely linked markers in two or more populations, although the markers represent only a small proportion of the polymorphisms in a short DNA sequence and the sampled populations are loosely defined (Table 2). Eaves *et al.* (15) reported 21 microsatellites in a 6.5-centimorgan interval on chromosome 18q21. Four populations were represented (Finland, Sardinia, United Kingdom, and United States), each with 800 unrelated haplotypes inferred from random families. Each locus was dichotomized for comparability with SNPs. Taillon-Miller *et al.* (16) studied 39 SNPs on male X chromosomes in 92 Europeans (from the Centre d'Etude du Polymorphisme Humain, Paris), 34 SNPs in 100 Finns, and a nonrandom sample of 17 SNPs in 150 Sardinians. Dunning *et al.* (17) reported small numbers of SNPs for 230 Afrikaners, 517 Ashkenazim, 432 Finns, and 376 East Anglican British in three short regions of 13q, 19q, and 22q. These studies provide enough information to fit the Malecot model for physical distance of markers shared by populations, but some intervals are too large to give reliable LD maps. Gabriel *et al.* (18) studied 3,738 SNPs in 54 autosomal regions representing 13.4 Mb. Four samples of unrelated individuals were examined: 67 Yorubans, 48 Europeans, 42 Chinese and Japanese, and 50 African Americans. We used National Center for Biotechnology Information BUILD 30 (June 2002) to

Table 4. Error variance (V) and relative efficiency (RE)

Source	Gabriel <i>et al.</i> (18)				Others (15–17)			
	–2ln lk	df	V	RE	–2ln lk	df	V	RE
Cosmopolitan map								
kb	261,915	251,150	1.043	0.686	7,502	2,051	3.658	0.559
LDU	205,719	248,336	0.828	0.863	4,617	1,975	2.337	0.875
Samples with cosmopolitan map								
ϵ estimated	189,792	248,031	0.765	0.934	4,449	1,949	2.282	0.896
ϵ, M estimated	186,551	247,787	0.753	0.950	4,212	1,927	2.186	0.936
Samples with own map	172,280	240,977	0.715	1.000	3,538	1,730	2.045	1.000

Table 5. Weighted difference estimates $\pm$ standard errors: Samples with cosmopolitan maps

Estimated	Parameter	European-Asian	African-American -Yoruban	African-Eurasian
ε estimated	ε	0.001 $\pm$ 0.043	0.162 $\pm$ 0.053	0.638 $\pm$ 0.046
ε, M estimated	ε	0.005 $\pm$ 0.048	0.110 $\pm$ 0.052	0.418 $\pm$ 0.052
ε, M estimated	$-\ln M$	-0.002 $\pm$ 0.006	0.026 $\pm$ 0.011	0.130 $\pm$ 0.013

make the physical maps for each dataset, rejecting 25 markers that could not be identified in the designated regions. In the absence of a finished map of the human genome, it is unprofitable to pursue discrepancies from the evolving draft sequence that may also implicate SNP names and primers.

Although ρ values are autocorrelated, their matrix is generally of full rank and gives the degrees of freedom in Table 3. Summing over samples, we determined error variances for alternative maps and efficiency relative to sample-specific LD maps (Table 4). The samples of Gabriel *et al.* (18) provide far more information (as relative efficiency) than the three studies at lower resolution, which were pooled. The cosmopolitan maps scaled to sample-specific values of ε with M estimated concurrently lose only $\approx 5\%$ of the information. Fixing the value of M from cosmopolitan maps is slightly less efficient. The cosmopolitan LDU maps themselves lose $>10\%$ of the information, whereas the kb map loses at least 30%. These results have far-reaching implications for positional cloning. Much effort is now being devoted to constructing a few LD maps at high resolution, partly in the hope that less informative SNPs may be eliminated in positional cloning. This would make sample-specific maps less efficient than in the data summarized here, where most markers were typed in all samples within a study contributing to a cosmopolitan map. Although sample size is relevant and the standard maps now under construction will be based on small samples, it is questionable whether a sample-specific map will have greater reliability than a cosmopolitan map scaled by estimating ε and M simultaneously. This substantially reduces costs of positional cloning, but requires care in choice of populations, individuals, and methods of map construction.

We next compared the populations sampled by Gabriel *et al.* (18) (Table 2), both separately and for differences within genomic region corresponding to African, others, and African vs. non-African. These results are concordant (Tables 5 and 6). The European and Asiatic samples, however examined, do not differ significantly in the ε and $-\ln M$ Malecot parameters. On the contrary, the African-American and Yoruban samples differ significantly in most comparisons, with African Americans consistently showing less LD as reflected by higher values of ε and $-\ln M$. The difference between African and non-African samples is highly significant, with less LD in the former as expected on the hypothesis that the diaspora of *Homo sapiens* from Africa to populate the rest of the world $\approx 100,000$ years ago was associated with a bottleneck that reduced genetic diversity and increased LD. There is no question that if only two cosmopolitan maps were chosen, they would correspond to the African and non-African dichotomy.

The lower resolution of a larger number of populations of European descent in other studies makes conclusions by stepwise

regression less clear (Table 7). A small decrease of $-\ln M$ in isolates is observed as expected, but like residual differences among populations does not approach significance. Estimates of ε are significantly lower for isolates than for regions, the ratio being about $(1.091 - 0.197)/1.091 = 0.82$. However, there is an enormous excess for the Afrikaner sample that disagrees with the historical record of a small founder population, increased frequency of a few major genes for diseases rare in other populations, and instances of persistent LD at large distance (19). Perhaps Afrikaner has both restrictive and loose applications. Finland has a suggestively lower ε than other isolates, whereas Sardinia is unremarkable. These results should be interpreted cautiously until large, well-characterized samples are compared with a cosmopolitan Eurasian standard. More extreme isolates defined by geographic or social barriers exist even within populations vaguely characterized as isolates and may be of interest if they provide enough cases of a particular disease for initial study and replication.

To summarize these results we excluded the Afrikaner sample as atypical of isolates and took Eurasian regions as a standard because they are the only population shared among studies. Estimating ε, M simultaneously we obtain the results in Table 8. African Americans have ε values about twice as great as European isolates, with corresponding expansion of their LD map. There are several sources of variation in these estimates, including choice of isolates and regional populations, sampling by appearance, residence, birthplace, parentage, self-identification, or presumed lack of admixture, selection of chromosome or segment, exclusion of small or variable allele frequencies, and adoption of a particular weight. In addition to the K_e weights presented here we tried kb and LDU weights. Estimates of scaling factors varied, but were not systematically different from Table 8. The impact of weights diminishes as sequence length increases. Selection of common SNPs (e.g., with a frequency of at least 0.1 in all populations) favors old polymorphisms that were not much affected by recent bottlenecks and therefore give inflated values of ε with less divergence among populations. There is little point in constructing many cosmopolitan maps, because each exercise in positional cloning will have its own scaling factor that may be estimated from its controls.

Discussion

Whereas there is no basis for choice between European descent and Asian samples, the African Americans representing many tribes have several advantages over a sample from a single tribe. Yorubans migrated in uncertain numbers to the lower reaches of the Niger river in the early 8th century AD. A relatively small founder population in the preceding century was likely, consistent with the unusually high frequency of dizygotic twinning in

Table 6. Weighted estimates $\pm$ standard errors: Samples with cosmopolitan map

Estimated	Parameter	Eurasian	Yoruban	African-American
ε	ε	0.597 $\pm$ 0.023	1.156 $\pm$ 0.044	1.388 $\pm$ 0.041
ε, M	ε	0.665 $\pm$ 0.024	1.065 $\pm$ 0.044	1.224 $\pm$ 0.038
ε, M	$-\ln M$	0.024 $\pm$ 0.004	0.123 $\pm$ 0.015	0.162 $\pm$ 0.014

Table 7. Weighted estimates for isolates and regions $\pm$ standard errors: Samples with cosmopolitan map

Estimated	Parameter	Region	Isolate – region	Afrikaner – region
ε	ε	1.071 $\pm$ 0.051	–0.205 $\pm$ 0.064	0.466 $\pm$ 0.159
	ε	1.071 $\pm$ 0.060	–0.178 $\pm$ 0.074	—
ε, M	ε	1.091 $\pm$ 0.051	–0.197 $\pm$ 0.064	0.452 $\pm$ 0.141
	ε	1.091 $\pm$ 0.062	–0.162 $\pm$ 0.077	—
	$-\ln M$	0.223 $\pm$ 0.049	–0.002 $\pm$ 0.043	—

the Yorubans, shared by their hundreds of gods and descendants in northeastern Brazil. There is strong evidence for one or more recessive genes predisposing to double ovulation, and therefore to dizygotic twinning (20). Whether LD in the Yorubans is typical of African tribes is doubtful, and they cannot represent the much larger population of sub-Saharan Africa. African Americans have less LD and are favorable for high-resolution positional cloning when non-African populations have failed to resolve predictive and causal SNPs. The aggregate population descended from African immigrants and now living in other countries (notably United States, Great Britain, Brazil, and countries in the Caribbean) is much larger than any single tribe resident in Africa. Each tribe may be comparable to a non-African isolate, but not to a large region. Individuals of African ancestry living outside of Africa share the environment of the larger community with similar life expectancies, health care, and disease incidences, making research in either group of general benefit. Science and the social contract speak in one voice for a cosmopolitan African LD map that is multitribal, like the large minority populations in developed countries.

The lower LD in Africa was noticed many years ago. Adam (21) reported a high frequency of repulsion between glucose-6-phosphate dehydrogenase (G6PD) deficiency and color blindness in Kurdish and Iraqi Jews, whereas Siniscalco *et al.* (22) found an excess of coupling in Sardinia. The two loci are closely linked on the X chromosome, and G6PD deficiency is protective against falciparum malaria. It is likely that G6PD deficiency was introduced into those isolates by few individuals, among whom color blindness was by chance at lower or higher frequency than in the general population. On the contrary, these loci are at equilibrium in African Americans (23), as expected for admixture of many tribes. This admixture is closer to Hardy–Weinberg equilibrium than any artificial mixture of samples from different tribes, and reliable methods are available to control for introgression from people of European descent.

The estimates in Table 6 provide some evidence about population history. The expected value of ε is proportional to θt , where θ is a small recombination fraction and t is the time since the last important bottleneck when population size was greatly reduced (8). For the non-Africans we assume that this was migration into Eurasia 100,000 years ago. It would be inappropriate to compare the Yoruban isolate with the Eurasian continent, and so the African Americans are a better yardstick. Assuming that θ is independent of ethnicity, Table 8 gives $(1.735)(100,000) = 173,500$ years as the time to the last major bottleneck in Africa. In the Malecot model ε is independent of

effective population size N_e and so a bottleneck in the African diaspora seems incontestable: the time is uncertain, but in any case roughly half the time to the last African bottleneck. Both estimates highly depend on assigning migration out of Africa to 100,000 years BP, but studies of mtDNA and Y chromosome DNA that do not make this assumption have yielded similar dates for the origin of modern humans consistent with their extermination of transitions between *Homo erectus* and *H. sapiens* (24, 25).

Assuming that M approached 1 for most human sequences at the first bottleneck and that the rate of mutation and gene conversion was much $< 1/2N_e$ per generation, where N_e is the harmonic mean of the effective population size per generation, we may estimate N_e as $t/-2\ln M$ with t measured in number of generations (8). Taking 173,500 years as 7,000 generations and $-\ln M$ as 0.162 from Table 6, $N_e = 21,600$. This result agrees with other estimates based on different assumptions (26). Weight assignment has a substantial effect when applied to many small sequences. The unweighted estimate is 0.224 ± 0.026 , giving $N_e = 15,600$. The effective size for Eurasia is less reliable because M in the diaspora from Africa probably did not approach 1 for most sequences. The unweighted mean of $-\ln M$ is 0.069 ± 0.011 . Assuming 100,000 years equal 4,000 generations, $N_e = 29,000$. Estimates of $-\ln M$ agree qualitatively with ε that migration from Africa was coincident with a bottleneck that reduced population size and diversity less than the primeval African bottleneck.

Population differences in LD pose unfamiliar problems. Confronted by samples from several ill-defined populations, we could not pool them without violating the Hardy–Weinberg assumptions on which estimation of haplotype frequencies from diplotypes currently depends. It is possible that a large sample from a single population will not be used as a standard, and that estimation of haplotype frequencies should be extended to replace the Hardy–Weinberg assumption by an inbreeding model (27), thereby increasing sample size and reducing the asymptote L , which is inflated when multiple samples are distinguished. In the interim we had no choice but to concatenate the samples and in this way take their association probabilities weighted by K_p as the data for construction of a standard LD map, against which the samples could be compared. This dilemma faces the HapMap project (28), which initially includes samples of roughly equal size from the three major racial groups, making the choice of any one of them for a standard map statistically inefficient, arbitrary, and contrary to prevailing notions of political correctness, although LD research is unlikely to be distributed equally over the three races. If, as currently proposed, additional populations are studied, not yet selected and perhaps with only a subset of markers, there may be no practical alternative to enlarging the potpourri to construct final standards. Our working hypothesis is that any population may be adequately represented by an LD map that differs from a cosmopolitan standard by Malecot parameters ε , M , and a sample-specific asymptote L that depends on the number of individuals and choice of allele frequencies. The populations compared here do not contradict this assumption. Instead of a

Table 8. Scaling factors for ε in LD maps

Table	European isolates	Eurasian regions	Yorubans	African Americans
5	—	1.000	1.463	1.629
6	—	1.000	1.602	1.841
7	0.819	1.000	—	—
Mean	0.819	1.000	1.532	1.735

single LD map, a cosmopolitan standard for each of the major racial groups would provide detailed comparison of LD patterns and give scaling factors closer to unity, therefore estimating high-resolution local maps more accurately. No sample from a local population is likely to be so large and reliably typed at such high resolution that its LD map would be as accurate as an appropriate cosmopolitan standard scaled to the local value of ϵ . If haplotype maps are ever defined (28), they will presumably annotate a small number of cosmopolitan LD maps from which genetic causes of disease may be efficiently localized by assigning a haplotype frequency vector to each individual. Because this

generates sample-specific frequencies that make cosmopolitan frequencies irrelevant, the information in a haplotype map is identical to the block-and-step structure revealed by an LD map, but excluding the additive LDU metric that facilitates positional cloning.

We are grateful to Alison Dunning, Iain Eaves, Stacey Gabriel, Patricia Taillon-Miller, and their colleagues for making their data available to us and Aravinda Chakravarti, James Crow, and Pui-Yan Kwok for helpful discussion. This work was supported by Grant G0100203 from the Medical Research Council.

1. Morton, N. E. (1973) in *Methods and Theories of Anthropological Genetics*, eds. Crawford, M. H. & Workman, P. L. (Univ. of New Mexico Press, Albuquerque), pp. 333–366.
2. Morton, N. E. (1982) in *Current Developments in Anthropological Genetics: Ecology and Population Structure*, eds. Crawford, M. H. & Mielke, J. H. (Plenum, New York), Vol. 2, pp. 449–466.
3. Shannon, C. E. (1948) *Bell System Tech. J.* **27**, 379–423 and 623–656.
4. Collins, A. & Morton, N. E. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 1741–1745.
5. Morton, N. E. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2556–2560.
6. Hill, W. G. & Robertson, A. (1968) *Theor. Appl. Genet.* **38**, 226–231.
7. Ohta, T. & Kimura, M. (1971) *Genetics* **68**, 571–580.
8. Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.-Y. & Collins, A. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 5217–5221.
9. Maniatis, N., Collins, A., Xu, C.-F., McCarthy, L. C., Hewitt, D. R., Tapper, W., Ennis, S., Ke, X. & Morton, N. E. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 2228–2233.
10. Goldstein, D. B. (2001) *Nat. Genet.* **29**, 109–111.
11. Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., Richter, D. J., Lander, E. S. & Altshuler, D. (2002) *Nat. Genet.* **32**, 135–142.
12. Zhang, W., Collins, A., Maniatis, N., Tapper, W. & Morton, N. E. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 17004–17007.
13. Devlin, B., Risch, N. & Roeder, K. (1996) *Genomics* **36**, 1–16.
14. Lonjou, C., Collins, A. & Morton, N. E. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 1621–1626.
15. Eaves, I. A., Merriman, T. R., Barber, R. A., Nutland, S., Tuomilehto-Wolf, E., Tuomilehto, J., Cucca, F. & Todd, J. A. (2000) *Nat. Genet.* **25**, 320–323.
16. Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P. & Kwok, P.-Y. (2000) *Nat. Genet.* **25**, 324–328.
17. Dunning, A. M., Durocher, F., Healey, C. S., Teare, M. D., McBride, S. E., Carlomagno, F., Xu, C.-F., Dawson, E., Rhodes, S., Ueda, S., *et al.* (2000) *Am. J. Hum. Genet.* **67**, 1544–1554.
18. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al.* (2002) *Science* **296**, 2225–2229.
19. Gordon, D., Simon, I. & Ott, J. (2000) *Genomics* **66**, 87–92.
20. Morton, N. E., Chung, C. S. & Mi, M.-P. (1967) *Genetics of Interracial Crosses in Hawaii* (Karger, Basel).
21. Adam, A. (1961) in *Proceedings of the Second International Congress on Human Genetics*, International Congress Series (Excerpta Medica and Istituto Gregorio Mendel, Rome), pp. 565–567.
22. Siniscalco, M., Bernini, B., Latte, B. & Motulsky, A. G. (1961) *Nature* **190**, 1179–1180.
23. Porter, I. H., Schulze, J. & McKusick, V. A. (1962) *Ann. Hum. Genet.* **26**, 107–122.
24. Hey, J. (1997) *Mol. Biol. Evol.* **14**, 166–172.
25. Agulnik, A. I., Zharkikh, A., Boettner-Tong, H., Bourgeron, T., McElreavey, K. & Bishop, C. E. (1998) *Hum. Mol. Genet.* **7**, 1371–1377.
26. Harpending, H. C., Batzer, M. A., Gurven, M., Jorde, L. B., Rogers, A. R. & Sherry, S. T. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 1961–1967.
27. Yasuda, N. (1968) *Biometrics* **24**, 915–935.
28. Couzin, J. (2002) *Science* **296**, 1391–1393.