Insights Into Recombination From Patterns of Linkage Disequilibrium in Humans

Susan E. Ptak,* Kristian Voelpel[†] and Molly Przeworski[†,1]

[†]*Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany and *Interdisciplinary Centre for Bioinformatics of the University of Leipzig, D-04103 Leipzig, Germany*

Manuscript received June 3, 2003
Accepted for publication January 19, 2004

ABSTRACT

An ability to predict levels of linkage disequilibrium (LD) between linked markers would facilitate the design of association studies and help to distinguish between evolutionary models. Unfortunately, levels of LD depend crucially on the rate of recombination, a parameter that is difficult to measure. In humans, rates of genetic exchange between markers megabases apart can be estimated from a comparison of genetic and physical maps; these large-scale estimates can then be interpolated to predict LD at smaller ("local") scales. However, if there is extensive small-scale heterogeneity, as has been recently proposed, local rates of recombination could differ substantially from those averaged over much larger distances. We test this hypothesis by estimating local recombination rates indirectly from patterns of LD in 84 genomic regions surveyed by the SeattleSNPs project in a sample of individuals of European descent and of African-Americans. We find that LD-based estimates are significantly positively correlated with map-based estimates. This implies that large-scale, average rates are informative about local rates of recombination. Conversely, although LD-based estimates are based on a number of simplifying assumptions, it appears that they capture considerable information about the underlying recombination rate or at least about the ordering of regions by recombination rate. Using LD-based estimators, we also find evidence for homologous gene conversion in patterns of polymorphism. However, as we demonstrate by simulation, inferences about gene conversion are unreliable, even with extensive data from homogeneous regions of the genome, and are confounded by genotyping error.

THE extent to which alleles assort randomly on chromosomes depends on the recombination rate, as well as on the demographic history of the species and on the selective pressures exercised on the genomic region. All else being equal, alleles at a given physical distance apart will tend to be more strongly associated—in greater linkage disequilibrium—when the recombination rate is lower. Thus, levels of linkage disequilibrium (LD) can be predicted from estimates of the recombination rate, if an adequate model for the history of the species exists (Ohta and Kimura 1971). Conversely, when an accurate estimate of the recombination rate is available, levels of LD carry considerable information about the history of the species and can therefore help to distinguish between alternative evolutionary models (*cf.* Wall 2001).

In humans, there is increasing interest in an accurate characterization of LD, prompted in part by the question of what marker density will be needed to attain reasonable power in genome-wide association studies (Kruglyak 1999; Gabriel *et al.* 2002; Phillips *et al.* 2003). The design of such studies would be facilitated by an ability to predict levels of LD in different population samples and regions of the genome. Such predictions depend crucially on recombination rate estimates. Short of typing markers in huge pedigrees, direct estimates of local recombination rates in humans are available only from sperm typing experiments (*e.g.*, Jeffreys *et al.* 2001). Unfortunately, these experiments are laborious and hence not feasible on a genomic scale. As a result, most efforts to predict the decay of allelic associations have had to rely on estimates of the recombination rate obtained from a comparison of physical and genetic maps (*e.g.*, Kruglyak 1999). These crude estimates, referred to hereafter as c_{map}, are based on the number of recombinants observed between markers megabases apart. They therefore need to be interpolated to predict LD at smaller scales. The reliability of the interpolation depends on the extent of local heterogeneity in the recombination rate.

To date, there is *direct* evidence for small-scale (<100 kb) variation in recombination rates for <10 regions of the genome (*e.g.*, Jeffreys *et al.* 2001; May *et al.* 2002). However, the observation of a block-like structure of LD, with long stretches of strong allelic associations followed by shorter segments with weak associations, has led researchers to suggest that there may be extensive recombination rate variation in the human genome (Daly *et al.* 2001; Gabriel *et al.* 2002). Although it

[1] *Corresponding author:* Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman St., Box G-W, Providence, RI 02912. E-mail: molly_przeworski@brown.edu

remains unclear how much rate variation, if any (Phillips *et al.* 2003), is needed to account for observed haplotype blocks, a recent study found that small-scale heterogeneity in recombination rates improved the fit of simple demographic models to LD data (Wall and Pritchard 2003a). These findings raise the possibility that c_{map} estimates will not be reliable predictors of local levels of LD (Stumpf and Goldstein 2003).

Recent studies of human patterns of LD have also highlighted the importance of a second feature of recombination: homologous gene conversion (Frisse *et al.* 2001; Przeworski and Wall 2001). Recombination is thought to result from the formation and resolution of Holliday structures. These resolutions can occur with exchange of flanking markers (crossover) or without (gene conversion alone; Griffiths *et al.* 1996). As pointed out by Andolfatto and Nordborg (1998), recombinants between markers megabases apart result overwhelmingly from crossover events, with a negligible contribution of gene conversion. Thus, c_{map} is essentially an estimate of the crossover rate alone. In contrast, recombinants between closely linked markers (*e.g.*, markers separated by 1 kb) result from both gene conversion and crossing over. If the odds of resolving a Holliday structure with or without a crossover are fixed throughout the genome, c_{map} estimates will underestimate local rates of genetic exchange (and hence overestimate levels of LD) by some set amount. If the odds vary, they may be highly inaccurate predictors. Either scenario may apply. Because it is difficult to study recombination between closely linked markers in mammals, almost nothing is known about gene conversion rates (Pittmann and Schimenti 1998). In summary, both homologous gene conversion and possible small-scale heterogeneity in the recombination rate cast doubt on the predictive value of c_{map}.

We tested how well c_{map} predicts local levels of LD in extensive polymorphism data collected by the SeattleSNP project (http://pga.mbt.washington.edu/). To do so, we characterized levels of LD by an estimate of the population rate of crossing over, $\rho = 4N_e c$ (N_e is the diploid effective population size and c the rate of crossing over per base pair per generation). This approach summarizes the LD in a region by a single number, so that levels of LD in different regions can be compared (Pritchard and Przeworski 2001). Further, $\hat{\rho}/4\hat{N}_e$ provides an estimate of the crossing-over rate that can be compared to c_{map} (Hudson 1987). We also estimated the ratio of gene conversion to crossover events from these same polymorphism data.

METHODS

SeattleSNP data: We used publicly available polymorphism data for autosomal genes implicated in inflammatory diseases (see Carlson *et al.* 2003). All the genic regions were resequenced in the same 24 African-American individuals and 23 individuals of European descent [from the Centre d'Etude du Polymorphisme Humain (CEPH)]. By resequenced, we mean that every base pair was called in every individual. However, at some positions, the genotypes in a subset of individuals could not be determined unambiguously, so they were considered to be missing (~4%; M. Rieder, personal communication). Since diploids were sequenced, the data consist of genotypes where the phase of multiple heterozygotes is unknown.

We analyzed the two groups separately because, in previous studies, allele frequencies and levels of LD differed between Sub-Saharan African and European populations (*e.g.*, Frisse *et al.* 2001). On the date of download (November 21, 2002), this represented 84 loci in African-Americans and 77 in the CEPH (see Table 1 of supplementary materials, available from http://email.eva.mpg.de/~przewors). We considered only data sets with >10 segregating sites, so that the median number of segregating sites is ~48 in African-Americans and 33 in European-Americans. In our usage, segregating sites include all biallelic markers, whether nucleotide substitutions or indels, but exclude a small fraction of sites with more than two alleles (18 across all regions). The regions always include a gene, but coding sequence represents a small proportion (~10% on average) of the total. The segments surveyed for variation in a region are not always contiguous: ~4–70 kb were sequenced (median of 11.2 kb) out of a total length of 4–180 kb (median of 11.8 kb).

Estimating the effective population size from diversity and divergence: Under the standard neutral model of a random-mating population of constant size, the diploid effective population size of humans, N_e, is the same for all regions of the genome (see Table 1 for a brief definition of symbols). It can be estimated as $\theta_W/(4\mu_{\text{div}})$, where μ_{div} is an estimate of the mutation rate per base pair per generation, μ, and θ_W is Watterson's estimate of the population mutation rate $\theta = 4N_e\mu$, based on the number of segregating sites in the sample (Watterson 1975). We estimated μ from human-chimpanzee divergence by concatenating all the loci, assuming 6 million years for the time to the common ancestor of a human and a chimpanzee sequence (Wall 2003) and a generation time of 20 years. This approach yielded $\mu_{\text{div}} = 1.69 \times 10^{-8}$/bp per generation. From the number of segregating sites across all loci, we obtained $\theta_W = 1 \times 10^{-3}$/bp for African-Americans and $\theta_W = 7 \times 10^{-4}$ for the CEPH. The estimate of N_e, N_{div} obtained in this way is ~15,000 for African-Americans and ~10,000 for the CEPH. Similar estimates are obtained by considering the mean or median N_e estimate across loci (results not shown).

The model of recombination: We considered a simple model of recombination to make inferences about rates of gene conversion (Wiuf and Hein 2000). The model assumes that recombination leads to one of two outcomes: either a small segment of DNA is copied from one chromosome to another (gene conversion alone)

TABLE 1

Symbols used

Symbol	Definition
c	The crossing-over rate per base pair per generation
c_{map}	An estimate of c obtained from a comparison of genetic and physical maps
μ	The mutation rate per base pair per generation
μ_{div}	An estimate of μ, based on human-chimp divergence
N_e	The diploid effective population size of humans
θ	The population mutation rate, $4N_e\mu$
θ_W	An estimate of θ
N_{div}	An estimate of N_e from diversity and divergence
g	Probability of initiation of a gene conversion event per base pair
f	$= g/c$
L	The mean length of a gene conversion tract
ρ	The population crossing-over rate, $4N_ec$
ρ_{LD}	An estimate of ρ based on patterns of linkage disequilibrium
ρ_{map}	$= 4N_{\text{div}}c_{\text{map}}$
ε	The genotyping error rate

or flanking markers are exchanged between the two chromosomes (crossing over). In other words, conversion and crossing over are treated as alternative resolutions of a Holliday structure and resolutions that yield a patchwork of DNA from both chromosomes are ignored. Gene conversion is unbiased, so each chromosome is equally likely to convert the other. Further, a gene conversion event is equally likely to initiate at any site, with probability g, and the length of the conversion tract is geometric, with mean L (Wiuf and Hein 2000). We denote the rate of crossing over per base pair per generation by c and define $f = g/c$ (following Frisse *et al.* 2001). In this model, the parameter f can be thought of as the odds of a Holliday junction resolving as a gene conversion *vs.* a crossover event.

Estimating the crossover rate: Rough estimates of the recombination rate can be obtained by a comparison of genetic and physical maps. We used sex-averaged recombination rate estimates based on the genetic map of Kong *et al.* (2002). The estimates of recombination rates represent average rates for a window of 3 Mb centered on a marker (Kong *et al.* 2002). To find the closest marker for each region, we repositioned the sequences on the August 2001 freeze and assigned each region to the closest marker; all but one locus (CCR2) was within 1.5 Mb of the marker. The estimates obtained in this way, denoted c_{map}, range from 0.23 to 3.32 cM/Mb; the mean is 1.20, close to the estimated genome average of 1.13 (Kong *et al.* 2002). Since for markers megabases apart the number of recombination events due to gene conversion is negligible, c_{map} is essentially an estimate of the crossing-over rate alone (Przeworski and Wall 2001). Thus, in what follows, c_{map} is referred to as an estimate of the crossover rate.

Estimating the population rate of crossing over when $f = 0$: We estimated the population rate of crossing over, $\rho = 4N_ec$, from polymorphism data using the method of Hudson (2001). We chose this method because simulations suggest that it performs as well or better than available alternatives that are computationally feasible for data sets of this size (Hudson 2001). The approach assumes the standard neutral model as well as an infinite-sites mutation model (where every mutation occurs at a new site). In the first analysis, we further assumed that there is no gene conversion, *i.e.*, that all Holliday structures are resolved with exchange of flanking markers alone. This is the model traditionally considered by population geneticists. Under these assumptions, levels of LD are a function of ρ and θ. Further, as θ approaches 0, and conditioning on two sites being polymorphic, the probability of a particular allelic configuration at a pair of sites is only a function of ρd, where d is the physical distance between the two sites (in base pairs).

Let $\mathbf{n}$ be the vector whose four entries are the counts of each haplotype in the sample. For a single pair, the maximum-likelihood estimate of ρ can be obtained by maximizing $\Pr(\mathbf{n}; \rho, d)$. To estimate ρ for polymorphism data from a given region of the genome, Hudson (2001) proceeds by maximizing the product $\Pi_i \Pr(\mathbf{n}_i; \rho, d_i)$ over all pairs i of sites in the region. This procedure ignores the dependence between pairs and hence the likelihood obtained is not a true likelihood, but instead is referred to as "composite likelihood." Note that because the allelic configurations depend on ρ, not c, we cannot estimate c and N_e separately.

We used diploid data, where the phase of double heterozygotes is unknown and in which data are missing; in addition, the ancestral state is unknown. Assuming random mating, the approach of Hudson (2001) can be modified accordingly, to obtain probabilities of diploid configurations conditional on ρ (Hudson 2001). Let $\mathbf{n_d}$ denote a vector with nine entries, with entries corresponding to the nine possible *distinguishable* diploid genotypes. The probability of $\mathbf{n_d}$ can be obtained from the probability of $\mathbf{n}$ by summing over the haploid configurations that could give rise to the diploid configuration, weighted appropriately; for details, see Hudson (2001). By maximizing the product $\Pi_i \Pr(\mathbf{n}_{\mathbf{d},i}; \rho, d_i)$, one can estimate ρ from data where the phase of double heterozygotes is unknown. The estimate of ρ obtained in this way is denoted ρ_{LD}. The program to estimate ρ from diploid data was kindly provided by R. Hudson. Simulations suggest that estimates of ρ based on n diploids are as accurate as, or slightly more accurate than, those based on n haploids but less accurate than estimates based on $2n$ haploids (S. E. Ptak, M. Przeworski and R. Hudson, unpublished results).

Likewise, the probabilities of allelic configurations where the ancestral state is unknown can be obtained from the probabilities when it is known, by summing over the appropriate sample configurations. Simulations suggest that there is not much loss of information when ancestral types are unknown (Hudson 2001). Configurations with missing data are dealt with similarly.

Estimating the population crossing-over rate when $f > 0$: We also considered a model where there is both gene conversion and crossing over. The probability of a configuration at a pair of sites, $\mathbf{n}_\mathbf{d}$, then depends on f, ρ, and d. Specifically, it depends on the rate at which a recombinant is produced between two sites, which is $r = c[d + 2fL(1 - \exp(-d/L))]$ under this model of recombination (Frisse *et al.* 2001). Assuming L is known, parameters f and ρ can be estimated by maximizing the product $\Pi_i \Pr(\mathbf{n}_\mathbf{d}; f, \rho, d_i)$. The program to do so was kindly provided by R. Hudson.

We estimated f and ρ using all the regions at once. Specifically, we estimated the likelihood of ρ and f values on a two-way grid, where ρ varies from 0 to 0.01 (in increments of 0.0001) and f from 0 to 10 (in increments of 0.25). We tried $L = 60$, 250, 500, and 1000, values consistent with the little that is known about gene conversion tract lengths in yeast, fruit flies, and humans (Frisse *et al.* 2001 and references therein). We did so under two sets of assumptions: In the first, which we refer to as the joint method, we assumed that f and ρ are the same across all loci and maximized $\text{CLik}(f, \rho) = \Pi_j \text{CLik}_j(f, \rho)$, where CLik_j is the composite likelihood for genic region j. We denote the composite maximum-likelihood estimate of ρ and f obtained in this way by $\hat{\rho}^+$ and $\hat{f}^+$, respectively. Frisse *et al.* (2001) used this method to analyze 10 short regions chosen to have similar recombination rates on the basis of large-scale c estimates.

For the SeattleSNPs data, c_{map} suggests that recombination rates vary by an order of magnitude across regions. We therefore developed a second approach, referred to as the profile method, by analogy to profile likelihood. This method assumes that f is fixed across loci, but allows ρ to vary. Let $\text{CLik}_j^{\text{P}}(f) = \text{CLik}_j(f, \hat{\rho}_f)$ be the profile composite likelihood of f for locus j, where $\hat{\rho}_f$ is the maximum composite-likelihood estimate of ρ for a given value of f. To obtain a maximum composite-likelihood estimate of f, $\hat{f}^*$, we maximized the product $\Pi_j \text{CLik}_j^{\text{P}}(f)$. For each locus, the estimate of ρ is given by $\hat{\rho}_{f^*}$.

Performance of estimators of f: To gain a rough sense of how well the two estimators of f are expected to perform on these data, we simulated 200 sets of loci that matched the actual data for African-Americans, with $f = 0, \ldots, 4$. For each locus, we generated data for 24 diploid individuals by coalescent simulations (Hudson 1990) of the standard neutral model (with an infinite-sites mutation model), setting $\rho = \rho_{\text{map}}$ and $\theta = \theta_{\text{W}}$. We then estimated ρ and f from each of the 200 sets of data using the joint and the profile method and tabulated how often we obtained particular values of $\hat{f}$ (on an integer grid from 0 to 8). In these simulations and all others described below, we matched the length of the sequence as well as the gaps for each region, but not the missing data, and created the observed number of genotypes by randomly pairing haplotypes (using modifications of software available from http://home.uchicago.edu/~rhudson1/source/mksamples.html).

Rejecting no variation in ρ across loci: We used coalescent simulations to test the null hypothesis of a fixed ρ across loci. Specifically, we considered the ratio of the maximum composite likelihood obtained under the null hypothesis and under the alternative model where f is fixed but ρ can vary [*i.e.*, $\text{CLik}(\hat{f}^+, \hat{\rho}^+)/\text{CLik}(\hat{f}^*, \hat{\rho}_{f^*})$]. (Note that when ρ is fixed across loci, the profile approach reduces to the joint one.) A small ratio suggests that the data are more probable under the alternative hypothesis that ρ is not fixed across loci. To assess significance, we compared the observed ratio to a distribution generated from 100 simulations with $\theta = \theta_{\text{W}}$, $\rho = \hat{\rho}^+$, and $f = \hat{f}^+$ for each locus.

Rejecting no gene conversion (*i.e.*, $f = 0$): We tested the null hypothesis that $f = 0$ in the CEPH and in African-American data. To do so, we compared the observed value of $\lambda = \text{CLik}(f = 0)/\text{CLik}(\hat{f}^*)$ to a distribution of λ values obtained from 100 simulations with $\theta = \theta_{\text{W}}$, $\rho = \hat{\rho}_0$, and $f = 0$ (*i.e.*, the estimate obtained from the profile method constraining f to be 0).

Effect of genotyping error on inferences about f: To examine the effect of genotyping error on estimates of f, we generated 100 sets of 84 simulated loci that matched the actual data for African-Americans, conditional on $\rho = \rho_{\text{map}}$. We set $f = 0$ but introduced "genotyping errors," then tabulated the proportion of sets where we obtained $\hat{f} > 0$ by either the joint or the profile estimation procedure. It is unclear how best to model genotyping error, since the process depends on the SNP detection technology and its use in specific cases. In addition, while rates of false positives can be estimated, rates of false negatives are much harder to assess. In our simulations, we chose an extremely simple model of genotyping error, meant to be illustrative rather than descriptive. To mimic a genotyping error rate of $\sim\varepsilon$, we switched each allele with probability $\varepsilon/2$ at every segregating site. As a result of these errors, some polymorphisms appear to be monomorphic, while others change frequency. Note that no additional, fake polymorphisms are created by this procedure, so that an implicit assumption is that the rate of false positive for low-frequency variants is 0; this may not be grossly unrealistic since sites with few heterozygous genotypes are often confirmed manually. This model also makes the sensible assumption that a heterozygote and a homozygote genotype are much more likely to be confused than are the two homozygote genotypes. We ran simulations for $\varepsilon = 0.005$ and 0.01.

To test whether we can reject the hypothesis of no gene conversion in the presence of genotyping error, we compared the observed λ to the distribution obtained from 100 simulations where $\rho = \hat{\rho}_0$, $f = 0$, and $\varepsilon = 0.005$.

Relationship between ρ_{LD} and c_{map}: To examine the relationship between ρ_{LD} and c_{map}, we estimated ρ_{LD} for each locus (for $f = 0, \ldots, 10$ in increments of 0.25). We then computed the rank correlation of these estimates and c_{map} estimates (for a given f), using Kendall's coefficient. To test whether this relationship would still be significant (at the 5% level) after correction for differences in N_e across regions, we considered the partial rank correlation of ρ_{LD} and c_{map} after correction for θ_W values. We estimated the significance of the observed partial correlation by a permutation test (with 100 permutations).

To gain a sense of our power to detect a relationship between ρ_{LD} and c_{map} using the SeattleSNPs data, we used coalescent simulations of the standard neutral model to generate 100 sets of 84 regions that mimicked the actual data. In each region, $\theta = \theta_W$ and $f = 0$. To take into account the uncertainty associated with estimates of c_{map}, ρ for each region was chosen from a gamma distribution with mean set to $\rho_{map} = 4N_{div}c_{map}$; the choice of gamma parameters was guided by the rough confidence intervals reported for the Kong *et al.* (2001) genetic map of humans (Weber 2002). We estimated ρ_{LD} from each set of simulated data, using the same approach as for the actual data. Using the 100 sets of simulated data, we then asked: (1) In what proportion of data sets is the correlation of ρ_{LD} and c_{map} (measured using Kendall's rank coefficient) as strong as or stronger than what we observed in the actual data?, (2) how often is it significant at the 5% level?, and (3) how often is the median ρ_{LD} as large as or larger than that observed? We also asked these questions using 100 sets of data generated with $f = 1$.

RESULTS

Levels of linkage disequilibrium in the African-American and CEPH samples: We analyzed 84 regions of the genome in a sample of 24 African-Americans and 77 regions in a sample of 23 CEPH individuals (see methods). For each region, we used θ_W (Watterson 1975) to estimate the population mutation rate, $\theta = 4N_e\mu$, and ρ_{LD} (Hudson 2001) to estimate the population crossover rate, $\rho = 4N_ec$, assuming no gene conversion (for the estimates of θ and ρ for each region, see Table 1 of the supplementary materials available at http://email.eva.mpg.de/~przewors). θ_W values in the two samples are highly correlated ($\tau = 0.527$, $p = 10^{-6}$, $n = 77$), as are ρ_{LD} values ($\tau = 0.541$, $p = 10^{-6}$, $n = 77$).

These estimators of θ and ρ assume the standard neutral model and it is unclear how to interpret the estimates under alternative models where N_e is not well defined. To compare samples, we therefore considered the ratio ρ_{LD}/θ_W, an estimate of the number of crossover events per mutation for each locus, c/μ (Andolfatto and Przeworski 2000; Fearnhead and Donnelly 2001). This ratio can be thought of as the number of crossover events per mutation needed to produce the observed levels of LD under the standard neutral model. Larger values reflect lower levels of LD and vice versa. By this approach, the median estimate of c/μ is 1.009 for the African-American sample and 0.451 for the CEPH one. Of the 77 regions where the comparison is possible, this value is larger for African-Americans in 60 cases, much more than would be expected by chance ($p = 10^{-6}$ by a two-tailed sign test). Thus, overall, there is more LD in the CEPH sample than in the African-American one.

More or less LD than expected? These LD-based estimates of c can be compared to large-scale estimates. The median c_{map}'s are 1.08 cM/Mb and 1.15 cM/Mb per base pair for the set of 84 loci in African-Americans and 77 loci in the CEPH sample, respectively. For the African-American sample, the median estimate of ρ_{LD} is 0.0010/bp. (Throughout, we consider the median ρ_{LD} and not the mean because with small probability the ρ estimator returns a very large value, so the mean is not well defined). Assuming a diploid effective population size of $N_{div} = 15{,}000$ for African-Americans (see methods), this yields a median crossover rate of $\rho_{LD}/4N_{div} =$ 1.71 cM/Mb. This value is ~58% larger than the median c_{map} estimate. In contrast, the median value of ρ_{LD} for the CEPH sample is 0.0003/bp. Assuming $N_{div} = 10{,}000$ for Europeans (see methods), this yields a median crossover rate of 0.75 cM/Mb, which is ~35% smaller than the median c_{map}.

To assess whether the discrepancies between c_{map} and LD-based estimates of the recombination rate are larger than expected by chance, we generated 100 sets of simulated data under the standard neutral model that mimicked the actual data. For each region, we set $\theta = \theta_W$ and chose ρ from a gamma distribution with mean $4N_{div}$ c_{map} (see methods). For the African-American sample, the median ρ_{LD} was equal to or larger than that observed in 0 of 100 cases, suggesting that there is significantly more recombination than expected from c_{map}. In other words, levels of linkage disequilibrium are unexpectedly low for the standard neutral model, when $f = 0$. The

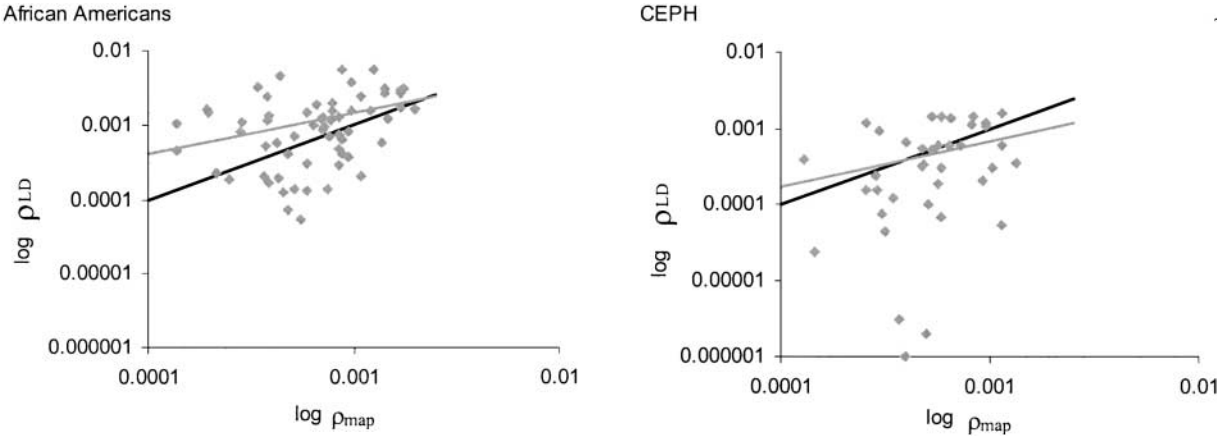

Figure 1.—Relationship between ρ_{LD} and $\rho_{map} = 4N_{div}c_{map}$ for data from African-American and CEPH samples. ρ_{LD} is estimated for no gene conversion (*i.e.*, $f = 0$). The solid line is what would be expected if $\rho_{LD} = \rho_{map}$ for all loci. The shaded line is Kendall's robust line fit (whose slope and intercept are the median slope and intercept of lines specified by all pairs of data points).

finding for the CEPH sample, however, is the opposite: Only in 1 of 100 cases was the median ρ_{LD} equal to or smaller than that observed. Thus, in the CEPH sample, levels of linkage disequilibrium are unexpectedly high for the standard neutral model, when $f = 0$.

These conclusions are predicated on an estimate of N_e, which in turn depends on the generation time and the time to a common ancestor, two parameters about which little is known (Wall 2003). In particular, if the generation time were >20 years (Helgason *et al.* 2003), the estimate of N_e would decrease. A smaller N_e (*e.g.*, 35% smaller) would explain the levels of LD seen in the CEPH sample but make the low levels of LD in the African-Americans even more extreme, while a larger estimate of N_e (*e.g.*, 58% larger) would have the opposite effect. Thus, while the discrepancies between observed and expected levels of LD are well within the range of our uncertainty about N_e, error in N_{div} alone cannot explain the patterns observed in both samples.

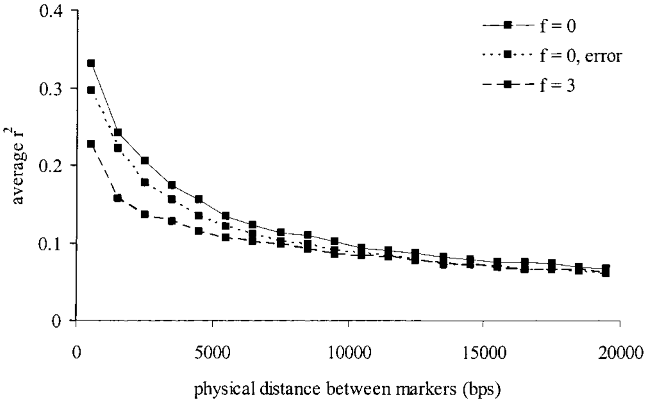

Figure 2.—The expected decay of r^2 with distance in a sample of 48 chromosomes for (1) a model with crossing over alone (solid line), (2) a model with crossing over and genotyping error (short-dashed line), and (3) a model with crossing over and gene conversion (long-dashed line). In all three, the population recombination rate $\rho = 0.001$/bp. In 2, the genotyping error rate $\varepsilon = 0.01$ (see methods), while in 3, the ratio of gene conversion to crossing over, $f = 3$. The expectation of r^2 was obtained from $>10^5$ coalescent simulations; each point represents the average r^2 for a 1000-bp interval.

Relationship of c_{map} and ρ_{LD}: To assess the extent to which average crossover rates between markers far apart predict local levels of LD, we considered the rank correlation of c_{map} and ρ_{LD} in both population samples. (We did not use parametric analyses, as many assumptions are grossly violated in this context.) When all loci with at least 10 segregating sites are considered, the two are significantly positively correlated at the 5% level in the CEPH sample ($\tau = 0.220$, $p = 0.007$, $n = 77$) but not in the African-American sample ($\tau = 0.116$, $p = 0.129$, $n = 84$).

Since estimates of ρ are highly inaccurate when there are few variable sites (Hudson 2001), we speculated that the lack of a significant relationship may be due to the inclusion of uninformative data sets. We therefore restricted our attention to data sets with 30 or more segregating sites. Although this cutoff decreases the number of loci we could consider, it ensures that ρ_{LD} estimates are within twofold of the true value of ρ ~90% of the time under the standard neutral model [Hudson's (2001) Figure 8, assuming $\theta = \rho$]. With this restricted data set, c_{map} and ρ_{LD} are significantly positively correlated in both population samples (see Figure 1; $\tau = 0.258$, $p = 0.025$, $n = 40$ for the CEPH sample and $\tau = 0.256$, $p = 0.004$, $n = 62$ for the African-Americans).

As reported previously, estimates of θ ($= 4N_e\mu$) are also positively correlated with c_{map} (Nachman 2001). Thus, estimates of ρ ($= 4N_ec$) could increase with c_{map} because of a relationship between N_e and c_{map}, as expected under models of variation-reducing selection (*cf.* Andolfatto 2001), rather than because of a relationship between small and large-scale recombination rates. However, in humans, the correlation between estimates of θ and c_{map} appears to be due to an association of mutation and recombination rates, not to a relationship between N_e and c_{map} (Hellmann *et al.* 2003). In addition, ρ_{LD} is still significantly correlated with c_{map} after correction for θ_W values ($\tau = 0.255$, $p = 0.03$, $n = 40$ for the CEPH sample and $\tau = 0.238$, $p = 0.01$, $n = 62$ for the African-Americans; see methods). Thus, it appears that large-scale estimates of the crossing-over rate, c_{map}, are informative about local levels of linkage disequilibrium, as measured by ρ_{LD}.

The strength of the underlying correlation between c_{map} and ρ_{LD} is unclear, however, as error in c_{map} and ρ_{LD} estimates will introduce noise and thus decrease the observed association. A related question is how often the observed correlation is expected under the standard neutral model, if $f = 0$ and in the absence of recombination rate heterogeneity. We examined this by tabulating how often a correlation is observed in simulated data that mimics the actual data (see methods). When all loci with at least 30 segregating sites were considered, there was a significant correlation between c_{map} and ρ_{LD} in 100 of 100 simulated data sets, for both population samples. Further, the rank correlation coefficient was always larger than that observed. When the cutoff was 10 segregating sites, the correlation was again always significant and only once was a rank correlation coefficient smaller than that observed (for the CEPH sample). [As expected, however, the correlation coefficients tended to be larger when we restricted our attention to only data sets with >30 segregating sites compared to when we considered all data sets with >10 (results not shown)]. In summary, if the assumptions of the model were met, a much stronger relationship would be expected between c_{map} and ρ_{LD} in spite of errors in both sets of estimates. This finding suggests that there are salient departures from model assumptions that reduce the correlation between large-scale and local recombination rates.

Performance of joint and profile estimators of *f*: As discussed above, one feature of recombination missing from the model of recombination is homologous gene conversion. To assess the evidence for gene conversion in the SeattleSNPs data, we used an extension of the method of Hudson (2001), implemented in Frisse *et al.* (2001). The observation behind this approach is that recombinants between closely linked sites result from both gene conversion and crossover events, whereas those between sites farther apart result almost exclusively from crossover events alone (Andolfatto and Nordborg 1998). (The exact definition of "close" depends on the distribution of gene conversion tract lengths.) As a result, there is a steeper decay of LD over short distances under a model with gene conversion and crossing over than under a model with only crossing over, while more distant markers show similar levels of associations under both models (as illustrated in Figure 2). The method that we used exploits the relationship between pairwise LD at different scales to coestimate ρ and f (for a given mean conversion tract length, L; see methods).

Unfortunately, when both ρ and f are estimated on a single locus, even large (*e.g.*, with >200 segregating sites), the power to reject $f = 0$ is low and estimates of the parameters are highly inaccurate (results not shown). We therefore assumed a fixed f value across loci and combined information from all loci to estimate this parameter. As described in methods, we did so under two sets of assumptions about ρ. In the so-called "joint" approach, we assumed that ρ is the same for all loci; in the second "profile" method, we allowed ρ to vary.

We tested the performance of both estimators of f by generating simulated data sets that mimicked the African-American data (see methods). For each region, we set $\theta = \theta_W$ and $\rho = \rho_{map}$ (for all values of f); thus, in our simulations, the population crossover rate varied across loci, but f was fixed. As can be seen in Figure 3, the joint estimator tends to overestimate the true value of f under these conditions. In contrast, the profile approach tends to underestimate it, but to a lesser extent. Similarly, the joint estimate of ρ is an underestimate, and the median profile estimate of ρ a slight overestimate, of the true median ρ (see Table 2 in supplementary materials at http://email.eva.mpg.de/~przewors). Overall, the profile estimator of f performs better than the joint method.

Which method is preferable in general depends on the extent to which ρ varies across regions. If it does not vary much, then the joint estimator should be more accurate, as it combines information from all regions to estimate ρ. In practice, researchers will rarely have precise estimates of the local ρ. At best, they will have a set of regions that are thought to experience similar crossing-over rates based on large-scale c estimates. To quantify the performance of the two estimators for these types of data, we generated 100 sets of 50 loci, where each locus was similar in information content to the data sets collected by Frisse *et al.* (2001). To allow for measurement error, we drew ρ values for each region independently from the same gamma distribution. As expected, in this situation, the joint method leads to more accurate estimates for $f > 0$ than does the profile approach. However, the point estimates of f are often >0 in the absence of gene conversion (~26% of the time; see Table 3 in supplementary materials at http://email.eva.mpg.de/~przewors).

Estimates of gene conversion rates: For the SeattleSNPs data, the c_{map} estimates point to substantial variation in ρ across regions. We therefore tested the null hypothesis that ρ and f are fixed across loci against an alternative where f is fixed but ρ can vary (see methods). We found that the polymorphism data are significantly more likely under the model where ρ is allowed to vary (the observed ratio of composite likelihoods was smaller than those for all 100 simulations). Consequently, we present estimates of ρ and f obtained using the profile approach. To facilitate comparison with previous studies, notably Frisse *et al.* (2001), we present results for $L =$

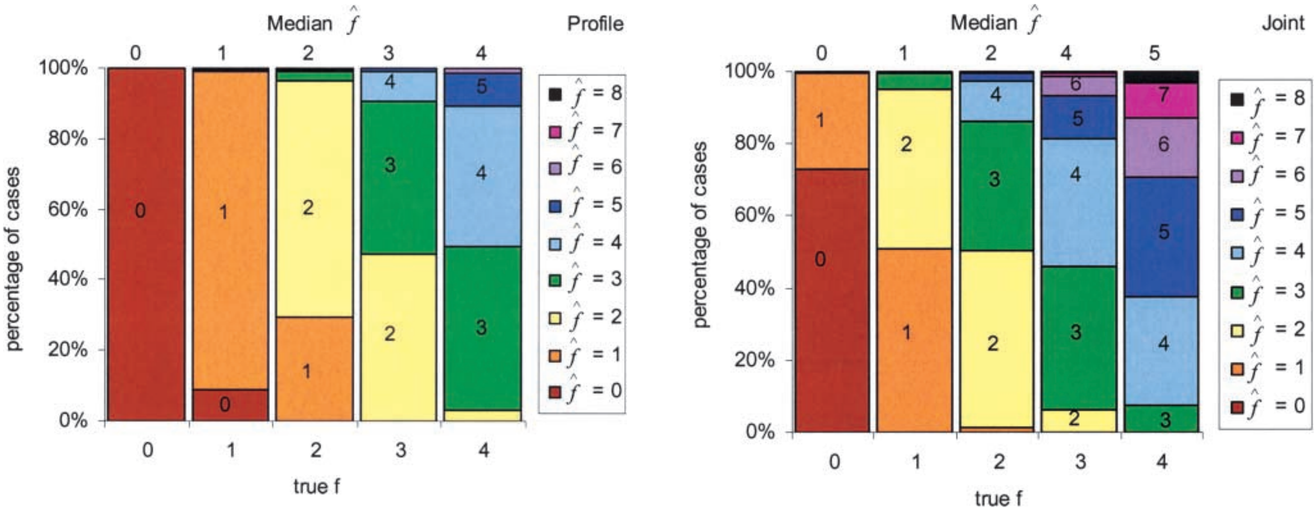

FIGURE 3.—The performance of estimators of *f*. Two hundred simulated sets of the 84 loci were generated for a given *f* value, conditional on $\rho = \rho_{map}$; the median ρ_{map} is 0.00065. For each locus, other parameters were chosen to mimic what is observed in the African-American sample. Shown here are the distributions of estimates of *f*, $\hat{f}$, obtained using either the profile or the joint estimation method (see METHODS for details), as well as the median $\hat{f}$.

500. The estimate of *f* increases with decreasing *L* (see Table 4 of supplementary materials at http://email.eva.mpg.de/~przewors).

For the African-American sample, the profile method yielded $\hat{f} = 1$ and median $\rho_{LD} = 0.0008$ while for the CEPH sample we obtained $\hat{f} = 0.25$ and median $\rho_{LD} = 0.0003$. Furthermore, simulations (see METHODS) suggest that the null hypothesis of no gene conversion can be rejected for both population samples (the observed λ is larger than that obtained in all 100 simulations for the African-American sample and in 96 out of 100 simulations for the European-American sample.

Effects of mutation rate variation on estimates of ρ and *f*: Inferences about ρ and *f* are potentially confounded by factors that affect allelic associations similarly. The estimator of ρ and *f* (HUDSON 2001*)* assumes an infinite-sites mutation model. It is known, however, that CpG dinucleotides mutate more frequently than do other pairs of sites (COOPER and KRAWCZAK 1989). This may lead to incorrect estimates of recombination parameters, as multiple hits to the same site can appear to be the result of recombination under the infinite-sites mutation model. To examine the potential effect of mutation rate variation, we estimated ρ and *f* for the African-American sample (with $L = 500$) after excluding all CpG sites from our polymorphism data (a CpG site was conservatively defined as any dinucleotide where at least one chromosome in the sample could have a CpG). Estimates were $\hat{f} = 1$ and median $\rho_{LD} = 0.0009$; thus, they were essentially unchanged by the exclusion of CpG sites.

Effect of crossover rate variation on inferences about *f*: An additional concern might be that the presence of short segments with elevated crossing-over rates ("recombination hotspots") can mimic the effects of gene conversion by decreasing LD between closely linked pairs. However, for the method of HUDSON (2001), this is not the case. Recall that the estimate of *f* is obtained by maximizing the product of the likelihood of *f* over all pairs of sites. If there are hotspots, closely linked pairs that span the hotspot exhibit lower levels of LD than they would in the absence of rate variation. However, most closely linked pairs do not span the hotspot and are therefore in stronger LD than expected. As a result, overall levels of LD at short distances are not lower than those in the absence of hotspots and variation in crossover rates does not lead to spurious inferences of gene conversion (results not shown).

Effect of genotyping error on inferences about *f*: A more serious problem is that genotyping errors have more effect on LD at short than at long distances and thereby mimic the effect of gene conversion (see Figure 2). To examine the effect of genotyping error on inferences about *f*, we ran simulations *with no gene conversion* but with genotyping error and estimated *f* (see METHODS). As can be seen in Figure 4, a 1% genotyping error rate led to a point estimate of $f > 0$ in all 100 simulated data sets, using either the profile or the joint approaches. Even with a seemingly small genotyping error rate (0.5%), $\hat{f} > 0$ in 78 and 99% of simulated runs, respectively. For the SeattleSNPs data, an error rate was assessed by genotyping a subset of the sites using two other genotyping technologies; ~0.5% of genotypes checked in this way differed from the original call (CARLSON *et al.* 2003; M. RIEDER, personal communication). Allowing for a 0.5% genotyping rate, we can no longer reject the hypothesis of no gene conversion under this (admittedly arbitrary) model of genotyping error (7 of 100 simulations have a smaller ratio than that observed; see METHODS). Thus, although gene conver-

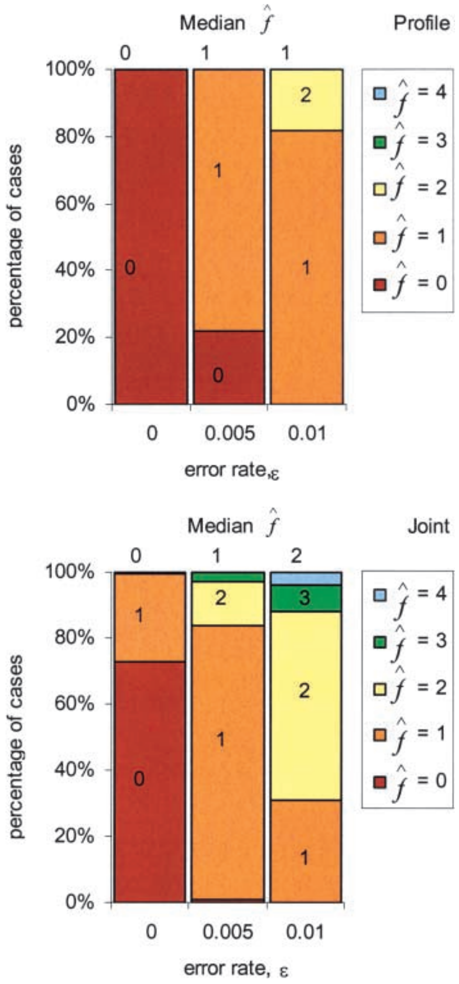

Figure 4.—Effect of genotyping error on inferences about *f*. One hundred simulated sets of 84 loci were generated as in Figure 3, but with no gene conversion (*i.e.*, $f = 0$). Genotyping errors were introduced at rate ε (see methods). Shown here is the distribution of estimates of *f*, $\hat{f}$, using either the profile or the joint estimation method (see methods) for three error rates, ε = 0, 0.01, and 0.005.

sion undoubtedly occurs in humans (Pittmann and Schimenti 1998; Jeffreys and Neumann 2002), we cannot distinguish its effects from genotyping error using the approach of Hudson (2001).

DISCUSSION

Too little LD in African-Americans and too much in the CEPH? In the SeattleSNPs data, levels of LD in the CEPH are higher than those in the African-Americans. This finding is consistent with reports of higher levels of LD in European populations relative to Sub-Saharan African ones (*e.g.*, Reich *et al.* 2001) and with the observation that European populations tend to have longer haplotype blocks (Gabriel *et al.* 2002; Wall and Pritchard 2003a). Because we summarized levels of LD by an estimate of the number of crossover events per mutation, c/μ, we could exclude a larger effective population size of African-Americans as an explanation. Indeed, the median estimate of $\theta = 4N_e\mu$ is ~1.6-fold higher in African-Americans, while the median estimate of $\rho = 4N_e c$ is >3-fold higher. Thus, levels of LD differ more between populations than do diversity levels, as previously reported for a comparison of Hausa and Italians (Frisse *et al.* 2001).

Our simulations further suggest that the levels of LD observed in African-Americans are lower than what would be expected under the standard neutral model (assuming $f = 0$). As discussed above, a trivial explanation is that we underestimated N_e. However, lower than expected levels of LD have also been reported in a study of 10 intergenic regions in a Hausa sample from Africa (Frisse *et al.* 2001) and of nine data sets collected in worldwide samples (Przeworski and Wall 2001), where N_e estimates were obtained under different assumptions. Furthermore, the simulations do not include gene conversion. Once this feature is incorporated, the standard neutral model appears to be an adequate model for levels of linkage disequilibrium in African-American samples (as measured by ρ_{LD}). As an illustration, when $f = 1$, levels of LD in African-Americans are closer to what is expected from a comparison of genetic and physical maps: The median crossing-over rate estimated from polymorphism data using the profile method, $\rho_{LD}/4N_{div}$, is 1.33 cM/Mb, while the median c_{map} is 1.08 cM/Mb. Moreover, the median ρ_{LD} is no longer significantly larger than expected from simulations where $f = 1$ and $\rho = \rho_{map}$ (results not shown).

Of course, a departure from model assumptions other than gene conversion could also have decreased levels of LD below the expectations of the standard neutral model, for example, population growth (*cf.* McVean 2002). It is worth noting, however, that most departures from model assumptions, including population structure and bottlenecks, as well as small-scale variation in recombination rates, will tend to have the opposite effect (Pritchard and Przeworski 2001; Wall and Pritchard 2003b). In particular, an obvious feature of African-American populations, population admixture, would be expected in theory to result in an increase of LD, thereby *decreasing* estimates of ρ (Chakraborty and Weiss 1988). In practice, levels of LD >10–100 kb appear quite similar in African-Americans and a Sub-Saharan African population (Wall and Pritchard 2003b). To summarize, it appears that there is less LD than expected in African-Americans under the simplest formulation of the standard neutral model; a plausible explanation is homologous gene conversion (or possibly genotyping errors).

In contrast to the African-American sample, the CEPH population shows unexpectedly high levels of linkage disequilibrium relative to the expectations of the standard neutral model when $f = 0$. When $f > 0$, the apparent excess of LD in the CEPH sample becomes more extreme (results not shown). This observation could be partially explained if N_e is *over*estimated. However, other features of polymorphism data in individuals of European descent suggest a demographic explanation may be more likely. In particular, the observation of reduced levels of diversity relative to Sub-Saharan African samples and an excess of intermediate frequency alleles in European samples have led to the suggestion of a recent reduction in population size in Europeans (Tishkoff *et al.* 1996; Frisse *et al.* 2001). A population bottleneck may also account for the relative excess of linkage disequilibrium (Tishkoff *et al.* 1996; Reich *et al.* 2001; Wall *et al.* 2002).

Inferences about gene conversion: Theoretical investigations have highlighted the potential importance of gene conversion in shaping local levels of linkage disequilibrium (Andolfatto and Nordborg 1998). Unfortunately, this process is difficult to study: Gene conversion events occur extremely rarely per base pair so that direct measurements often require the examination of a prohibitive number of meioses. An alternative approach, employed here, is to estimate gene conversion rates from polymorphism data. We estimated the ratio of gene conversion to crossover events, *f*, to be ~1 in the African-Americans and 0.25 in the CEPH. Simulations suggest that both estimates are significantly >0, assuming no or very little genotyping error. These represent the second estimates of *f* from polymorphism data. In the first, the estimate was based on a smaller data set sequenced in a Sub-Saharan African sample and obtained under the assumption that crossing-over rates are fixed across loci (Frisse *et al.* 2001). The point estimate, *f*, was substantially higher than ours; the discrepancy may reflect an upward bias in the estimator, as our simulations suggest that the joint estimate of *f* and ρ overestimates the true *f* when rates vary substantially across loci (Figure 3).

These inferences about *f* rely on a number of assumptions. Most obviously, they are based on a simplified model of recombination, for which there is support in humans from single-sperm typing (*e.g.*, Jeffreys *et al.* 2000), but about which much remains unknown. Furthermore, because of computational limitations, estimates of *f* condition on a particular value of the average gene conversion tract length, *L*. In particular, we have presented results for $L = 500$ bp to facilitate the comparison with Frisse *et al.* (2001). Since the estimates of *f* increase with decreasing *L* (Table 4 of supplementary materials at http://email.eva.mpg.de/~przewors), rates of gene conversion may be much higher than those reported if the average length of a gene conversion tract is ≪500 bp. Third, because estimates of *f* based on data from a single locus are extremely inaccurate, we (and others) have assumed that the ratio of gene conversion to crossing over is fixed across loci, an assumption that may be invalid (see below). Finally, our simulations (Figures 2 and 4) suggest that inferences about *f* are highly sensitive to even small levels of genotyping errors. Some of these difficulties may be overcome by the development of multilocus approaches to estimate gene conversion, such as the extension of the method of Hudson (2001) to consider triplets of sites rather than pairs (J. D. Wall, personal communication). It would also be useful to have more resequencing data such as these from regions with well-estimated crossing-over rates and for other populations. Finally, single-sperm typing experiments designed to estimate rates of homologous gene conversion would help to specify *L* and *f* appropriately.

Contrasting local and large-scale estimates of the recombination rate: Local recombination rates estimated from polymorphism data (ρ_{LD}) increase with large-scale estimates of the crossing-over rate (c_{map}) in both African-Americans and the CEPH. This association suggests that LD-based estimates of the recombination rate such as ρ_{LD} capture considerable information about underlying recombination rates or at least about the ordering of different regions by levels of LD—in spite of their reliance on a number of unrealistic assumptions, such as a constant population size and random mating. Consistent with our conclusion, recent studies found close agreement between LD-based estimates of recombination rates (using a different approach) and single-sperm typing estimates at the TAP2 and β-globin loci in humans (Li and Stephens 2003; Wall *et al.* 2003). Thus, it appears that polymorphism-based estimates of the recombination rate represent a useful tool for characterizing local levels of linkage disequilibrium.

While large-scale estimates of the crossing-over rate are predictive of local levels of LD, simulations suggest that the association of c_{map} and ρ_{LD} is weaker than would be expected if there were no heterogeneity in recombination rate or variation in *f* across loci (see results). One explanation is that *f* varies substantially across loci, reducing the strength of the relationship. To test this possibility, we ran 100 simulations where *f* was not fixed across loci, but was instead an integer "chosen uniformly" between 0 and 9. The association did tend to be weaker in this situation compared to one where *f* was fixed and nonzero across regions (results not shown). However, the observed correlation coefficient between c_{map} and ρ_{LD} (estimated for any fixed $f \leq 9$) was smaller than that seen in all 100 simulations (results not shown). Thus, variation in *f* alone is unlikely to explain the weakness of the correlation. A plausible alternative is that occasional recombination hotspots are reflected in the SeattleSNPs data. Candidates for hotspot regions are the loci with unusually high ρ_{LD} estimates given c_{map} (see Figure 1).

Outlook: We find that average recombination rates over large distances are informative about local rates of genetic exchange and, hence, about local patterns of linkage disequilibrium. This finding is somewhat surprising: In light of increasing evidence for extensive local variation in recombination rates, why do regions of 4–180 kb so often reflect the rates obtained from averaging over megabases? One possibility is that large-scale rates reflect the background rate of recombination (*i.e.*, the rate outside of hotspots). To address this and related questions, further work is needed to quantify how much recombination occurs in hotspots and the variation in intensity among hotspots, as well as to assess the extent to which global features of chromosome structure influence the location of recombination events (Petes 2001; De Massy 2003).

We thank D. Cutler, R. Hudson, M. Stephens, and J. Wall for useful discussions and P. Andolfatto, Y. Gilad, J. Wall, and three anonymous reviewers for comments on the manuscript. We are also grateful to I. Hellmann and N. Matasci for bioinformatics help; to M. Rieder for providing additional information about the SeattleSNP data; and especially to D. Nickerson, M. Rieder, and the SeattleSNPs project for making their data publicly available. This work was supported by Deutsche Forschungsgemeinschaft grant BIZ6-1/1.

LITERATURE CITED

Andolfatto, P., 2001 Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev. **11:** 635–641.

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397–1399.

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

Carlson, C. S., M. A. Eberle, M. J. Rieder, J. D. Smith, L. Kruglyak *et al.*, 2003 Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat. Genet. **33:** 518–521.

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

Cooper, D. N., and M. Krawczak, 1989 Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. **83:** 181–188.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander, 2001 High-resolution haplotype structure in the human genome. Nat. Genet. **29:** 229–232.

De Massy, B., 2003 Distribution of meiotic recombination sites. Trends Genet. **19:** 514–522.

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

Griffiths, A. J. F., J. H. Miller, D. T. Suzuki, R. C. Lewontin and W. M. Gelbart, 1996 *An Introduction to Genetic Analysis*. W. H. Freeman, New York.

Helgason, A., B. Hrafnkelsson, J. R. Gulcher, R. Ward and K. Stefansson, 2003 A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. Am. J. Hum. Genet. **72:** 1370–1388.

Hellmann, I., I. Ebersberger, S. Ptak, S. Paabo and M. Przeworski, 2003 A neutral explanation for the correlation of diversity with recombination in humans. Am. J. Hum. Genet. **72:** 1527–1535.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., 1990 *Oxford Surveys in Evolutionary Biology*, pp. 1–44. Oxford University Press, Oxford.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

Jeffreys, A. J., and R. Neumann, 2002 Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat. Genet. **31:** 267–271.

Jeffreys, A. J., A. Ritchie and R. Neumann, 2000 High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum. Mol. Genet. **9:** 725–733.

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

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

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

Li, N., and M. Stephens, 2003 Modelling linkage disequilibrium and identifying recombination hotspots using SNP data. Genetics **165:** 2213–2233.

May, C. A., A. C. Shone, L. Kalaydjieva, A. Sajantila and A. J. Jeffreys, 2002 Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. Nat. Genet. **31:** 272–275.

McVean, G. A., 2002 A genealogical interpretation of linkage disequilibrium. Genetics **162:** 987–991.

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17:** 481–485.

Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68:** 571–580.

Petes, T. D., 2001 Meiotic recombination hot spots and cold spots. Nat. Rev. Genet. **2:** 360–369.

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding *et al.*, 2003 Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat. Genet. **33:** 382–387.

Pittmann, D. L., and J. C. Schimenti, 1998 Recombination in the mammalian germ line. Curr. Top. Dev. Biol. **7:** 1–35.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.

Przeworski, M., and J. D. Wall, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet. Res. **77:** 143–151.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Stumpf, M. P., and D. B. Goldstein, 2003 Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr. Biol. **13:** 1–8.

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

Wall, J. D., 2001 Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. Curr. Opin. Genet. Dev. **11:** 647–651.

Wall, J. D., 2003 Estimating ancestral population sizes and divergence times. Genetics **163:** 395–404.

Wall, J. D., and J. K. Pritchard, 2003a Assessing the performance of the haplotype block model of linkage disequilibrium. Am. J. Hum. Genet. **73:** 502–515.

Wall, J. D., and J. K. Pritchard, 2003b Haplotype blocks and the structure of linkage disequilibrium in the human genome. Nat. Genet. Rev. **4:** 587–597.

Wall, J. D., P. Andolfatto and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203–216.

Wall, J. D., L. A. Frisse, R. R. Hudson and A. D. Rienzo, 2003 Comparative linkage disequilibrium analysis of the β-globin hotspot in primates. Am. J. Hum. Genet. **73:** 1330–1340.

Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.

Weber, J. L., 2002 The Iceland map. Nat. Genet. **31:** 225–226.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Communicating editor: J. J. Hein