Genet. Res., Camb. (2001), **77**, *pp.* 143–151. Printed in the United Kingdom © 2001 Cambridge University Press

Why is there so little intragenic linkage disequilibrium in humans?

MOLLY PRZEWORSKI[1] AND JEFFREY D. WALL[2]*
[1] *Statistics Department, Oxford University, 1 South Parks Road, Oxford OX1 3TG, UK*
[2] *2102 Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA*

(*Received 12 April 2000 and in revised form 16 October and 28 November 2000*)

Summary

The efficient design of association mapping studies relies on a knowledge of the rate of decay of linkage disequilibrium with distance. This rate depends on the population recombination rate, C. An estimate of C for humans is usually obtained from a comparison of physical and genetic maps, assuming an effective population size of approximately 10^4. We demonstrate that under both a constant population size model and a model of long-term exponential growth, there is evidence for more recombination in polymorphism data than is expected from this estimate. An important contribution of gene conversion to meiotic recombination helps to explain our observation, but does not appear to be sufficient. The occurrence of multiple hits at CpG sites and the presence of population structure are not explanations.

1. Introduction

A major goal of human genetics is the identification of loci that contribute to the risk of non-Mendelian diseases. Theoretical studies have demonstrated that association mapping may be a powerful approach if the variants typed include the disease-susceptibility locus itself (Risch & Merikangas, 1996). However, short of genotyping all variable markers in the genome, association studies may not include the disease-associated site. The aim is then to find a single nucleotide polymorphism (SNP) in linkage disequilibrium with the susceptibility allele. The power of such association studies will depend on the regional patterns of linkage disequilibrium (Long & Langley, 1999). The efficient design of association mapping studies will therefore be aided by a knowledge of factors influencing linkage disequilibrium (cf. Kruglyak, 1999).

Linkage disequilibrium (LD) is shaped by numerous forces including genetic drift, natural selection and the population recombination rate. Consider, for example, the model commonly used to estimate population parameters: a randomly mating population at mutation–drift equilibrium (Kimura, 1968). Under this set of assumptions, and an infinite-sites mutational model, the mean squared correlation between sites is given by $\frac{1}{(1+4N_e r)}$, where N_e is the diploid effective population size of the species and r the per generation rate of recombination between sites (Ohta & Kimura, 1971).

Thus, a key parameter in the design of association studies is the *population* recombination rate, $C = 4N_e r$ (Long & Langley, 1999). C can be estimated in two ways: In the first, r is estimated from a comparison of genetic and physical maps while N_e is estimated from levels of diversity (e.g., Kruglyak, 1999); this method relies on the standard neutral model outlined above and an estimate of the mutation rate per generation. In what follows, we refer to this estimate of C as C_{map}. Alternatively, r and N_e can be estimated jointly, from nucleotide sequence polymorphism data (e.g. Hudson, 1987; Griffiths & Marjoram, 1996; Wall, 2000). In this approach, the observed patterns of linkage disequilibrium are used to estimate C. If the assumptions behind these methods are realistic, the two estimation methods should yield similar results (Hudson, 1987; Andolfatto & Przeworski, 2000).

Here, we present a comparison of the two methods of estimation of C for the nine suitable human data sets currently available. We highlight a systematic discrepancy between these two methods for both a model of constant population size and one of long-

* Corresponding author. Tel: +1 (617) 495 1568. Fax: +1 (617) 496 5854. e-mail: jwall@oeb.harvard.edu

https://doi.org/10.1017/S0016672301004967

term exponential growth. There appears to be more evidence for recombination in polymorphism data than would be expected from C_{map}. A similar pattern has been found in other data (K. Ardlie & L. Kruglyak, personal communication).

2. Methods

We examine all nuclear sequence data sets for which frequency and haplotype data were available, the sample size (n) was at least 10, and the number of segregating sites (S) was at least 10. Only biallelic polymorphisms (both point mutations and indels) are included. We exclude mutations that overlap with deletions, since these offer incomplete information. For data sets other than *Zfx*, each individual is sequenced for the whole length (rather than at preselected variants); the data for *Zfx* was obtained by SSCP (for details see Przeworski *et al*., 2000). Haplotypes were determined by the method of Clark (1990) combined with allele-specific PCR (see Clark *et al*. 1998) for *Ace*, *ApoE* and *Lpl*. For the two other autosomal loci, they were determined experimentally (*β-globin*, *Duffy* gene sequence).

We use the largest world-wide sample available for each locus, even though some loci show substantial differentiation between populations. In particular, the sub-Saharan sample at *Duffy* locus is fixed for an allele that appears to be absent from other populations. Since the allele confers resistance to *Plasmodium vivax*, the African pattern of polymorphism is thought to reflect the action of natural selection (Hamblin & Di Rienzo, 2000). If we use only the Italian sample to estimate C, C_{HRM} increases (results not shown), so using a world-wide sample is conservative for our purposes.

Rates of recombination, r, are estimated from a comparison of the Genethon genetic map (Dib *et al*., 1996) and the GB4 radiation hybrid map (Gyapay *et al*., 1996). To estimate the regional recombination rate, we use the ratio of the genetic distance to the physical distance between the two (high confidence) microsatellite markers closest to the locus of interest. This method gives us an estimate of the number of cM/cR. To obtain an estimate of the number of cM/Mb, we use chromosome-specific conversion factors for cR to Mb (Hudson *et al*., 1995). Recombination rates are often estimated using more than two microsatellite markers to reduce the sampling variance associated with a small number of meioses. However, if rates vary on the scale of a couple of megabases, rates averaged over longer distances may not be informative locally. In addition, random error in r estimation will not lead to a systematic trend. Thus, for the purposes of this study, we choose to use only two markers. Estimates using other methods are listed in Table 1.

To obtain an estimate C_{map} of the population recombination rate C, we multiply r by $4N_e$ (or $2N_e$ if X-linked). N_e is assumed to be 10^4 (Li & Sadler, 1991; Takahata, 1993). This estimate of N_e is obtained from observed levels of diversity; it relies on assumptions of a panmictic population of constant size with no selection, and requires an estimate of the mutation rate per generation (see Section 3 for more details).

Estimates of C based on polymorphism data have high variances and are often biased (Wall, 2000). We chose an estimate (referred to as C_{HRM}) that appears to be roughly unbiased under the standard neutral model outlined above, has relatively low mean squared error, and can be calculated for large polymorphism data sets (Wall, 2000; J.D.W., unpublished results). C_{HRM} is a maximum likelihood estimator of C based on two summaries of the data: R_M, the minimum number of recombination events (cf. Hudson & Kaplan, 1985), and H, the observed number of distinct haplotypes (for more detail see Wall, 2000). The values of H may be underestimates for *Lpl* and *Ace*, due to incomplete phase information. For *Ace*, the presence of singletons on individuals *C08*, *C09* and *C24* (Rieder *et al*., 1999) increased the minimum number of inferred haplotypes to $H = 16$. The likelihood L $(C|R_M, H)$ is estimated from coalescent simulations (Kingman, 1982; Hudson, 1990). At least 2×10^5 replicates are run for each parameter combination. To determine credibility intervals for the maximum likelihood estimates, we employ the standard χ^2 approximation for the likelihood ratio statistic $-2\ln(L_1/L_0)$, where L_0 is the maximum likelihood and L_1 is the likelihood at an alternative point. We caution that there is no direct evidence that the standard χ^2 approximation is appropriate; future work will address this issue. Loci are assumed to be independent, so the likelihood of the whole data equals the product of the likelihoods at each locus. If the estimated likelihood for a particular locus is 0, we replace it with the reciprocal of the number of trials run to keep all calculations well defined. Most simulations were run using modifications of programs kindly provided by R. Hudson.

We assume a neutral infinite-sites model for our simulations. Our approach differs slightly from standard coalescent simulations: instead of conditioning on a population mutation rate, we generate genealogies and place the observed number of mutations, S, on the tree (cf. Hudson, 1993). Standard full likelihood methods exist that use all the information in the data (e.g., Griffiths & Marjoram, 1996), but they are computationally prohibitive.

Besides the standard, equilibrium panmictic null model, we consider two alternative demographic models. To incorporate recent population growth, we consider a model with a constant population size followed by exponential growth (cf. Marjoram &

https://doi.org/10.1017/S0016672301004967

Table 1. *Conflicting estimates of the population recombination rate in humans*

Locus	n^a	Bps	S^b	C_{HRM}	C_{map}	Other C_{map} estimates	Obs. R_M	P^c	References
β-globin	349	2670	19	24·0	2·4	2·0[d], 3·1[e]	3	0·039	Harding *et al.* (1997)
Duffy	82	1931	16	5·0	0·5	1·0[d]	2	0·010	Hamblin & Di Rienzo (2000)
Lpl	142	9700	87	120·0	9·3	4·2[f], 4·5[g], 8·8[d]	22	$< 10^{-5}$	Clark *et al.* (1998)
Ace	22	24000	78	17·0	9·1	1·6[e], 7·6[d]	6	0·153	Rieder *et al.* (1999)
ApoE	192	5491	22	23·0	6·5	3·3[f]	8	$< 10^{-5}$	Fullerton *et al.* (2000)
Dmd44	41	3000	17	60·0	2·1	0·1[e], 0·8[g], 2[d], 3·6[h], 4·6[i]	7	$< 10^{-5}$	Nachman & Crowell (2000*a*)
Pdha1	35	4200	24	6·2	5·0	0·2[e], 0·8[g], 3·2[i], 3·7[d]	3	0·148	Harris & Hey (1999)
Xq13·3	70	10163	33	2·0	0·3	0·3[j], 1·7[d]	1	0·239	Kaessmann *et al.* (1999)
Zfx	336	1089	10	3·8	0·1	0·04[e], 0·7[g], 0·8[d]	1	0·030	Jaruzelska *et al.* (1999)

[a] Number of chromosomes sequenced.
[b] Number of segregating sites.
[c] $P = \Pr(R_M \geqslant \text{obs. } R_M | C_{map})$.
[d] B. Payseur, personal communication. Rates are obtained from a comparison of the Genethon genetic map with the GB4 radiation hybrid map, using a sliding window approach of 5 microsatellites on each side of the locus of interest and cR-Mb chromosome-specific conversion factors from Hudson *et al.* (1995). A linear function was fit to estimate recombination rates. For details, see Payseur and Nachman (2000).
[e] Huttley *et al.* (1999).
[f] S. M. Fullerton, personal communication. Per band estimates of recombination rate were obtained from Morton's integrated database (Collins *et al.* 1996), averaged across three-band windows.
[g] B. Payseur, personal communication. Rates are estimated from the Morton integrated maps as in [e].
[h] Nachman & Crowell (2000*a*).
[i] Nachman *et al.* (1998).
[j] Kaessman *et al.* (1999).

Table 2. *Estimates of the population recombination rate under a model of exponential growth*

Locus	C_{map} [a]	Constant size C_{HRM}	Exp. growth[b] C_{HRM}	Exp. growth[c] C_{HRM}	P[d]
β-globin	2·4	24·0	42·0	24·0	0·051
Duffy	0·5	5·0	1·7	1·9	0·016
Lpl	9·3	120·0	—[e]	—[e]	$< 10^{-6}$
Ace	9·1	17·0	6·3	8·4	0·291
ApoE	6·5	23·0	—[e]	7·0	0·00003
Dmd44	2·1	60·0	35·0	27·0	$< 10^{-6}$
Pdha1	5·0	6·2	2·1	2·7	0·237
Xq13·3	0·3	2·0	0·5	0·7	0·334
Zfx	0·1	3·8	1·2	1·6	0·030

[a] C_{map} was estimated under the standard neutral model, so is a slight overestimate of $4N_0r$.
[b] The onset of growth is 50 Kya, and the population size increases from 10^4 to 10^6.
[c] The onset of growth is 50 Kya, and the population size increases from $N_0 = 10^4$ to $N_1 = 10^5$.
The C_{HRM} values reported are estimates of $4N_0r$.
[d] $P = \Pr(R_M \geqslant \text{obs. } R_M | C_{map})$. The model of growth is as in [c].
[e] Cannot be calculated because the estimated likelihood of the data is 0 (see text).

Donnelly, 1994). The population size increases from $N_0 = 10^4$ to $N_1 = 10^5$ or 10^6, where N_0 is the effective population size 50 thousand years ago (50 Kya), and N_1 is the current effective population size. The C_{HRM} values reported in Table 2 are estimates of $4N_0r$. To model population differentiation between African and non-African samples, we also run coalescent simulations of a symmetric two-island model (cf. Wright, 1931). The number of individuals drawn from each island corresponded to the actual sampling scheme for each locus. Migration rates were chosen to yield mean F_{ST} values (Wright, 1951) of 0·15 and 0·30. Higher values of F_{ST} correspond to more population differentiation. Most observed F_{ST} values are lower than 0·15 (see, e.g., Cavalli-Sforza *et al.*, 1994). Other aspects of the simulations were as above.

To test for the possible effects of multiple mutations at CpG sites, we rerun our likelihood simulations excluding all segregating sites where either allele forms a cg with an adjacent nucleotide site. This analysis is conservative, and excludes all transitions and transversions both to and away from CpG sites.

Finally, we consider an alternative model of recombination that incorporates both crossing-over and gene conversion. We generalize the standard coalescent with recombination (Hudson, 1990) by assuming gene conversion events (with geometrically distributed tract lengths) occur at constant rate on all the branches (cf. Wiuf & Hein, 2000). Since very little is known about gene conversion in mammals, we consider a model that is plausible based on research in *Drosophila* and yeast (e.g. Orr-Weaver & Szostak, 1985; Petes *et al.*, 1991; Hilliker *et al.*, 1994). We consider a mean tract length of 500 base pairs, and take the rate at which conversion events originate at a given base pair to be equal to the rate at which crossing over events occur at that base pair. In the text, C_{HRM} refers to the estimated population crossing-over rate. A C program written by the authors that incorporates gene conversion is available on request.

3. Results and discussion

As can be seen in Table 1, C_{HRM} is greater than C_{map} for 9 of 9 loci, for many loci by an order of magnitude. If either ordering of estimates were equally likely, the probability that all 9 have the same one by chance is $P = 0{\cdot}0039$ (two-tailed). In other words, the patterns of LD suggest more recombination than does an integration of genetic and physical maps. As an illustration, when 10^5 coalescent simulations are run for *Lpl* with $n = 142$, $S = 87$ and $C = C_{map}$, the simulated R_M value is *always* less than the actual R_M. Since the expected R_M increases with increasing recombination rate, it is likely that $C \gg C_{map}$ for *Lpl*. There is a similar situation for *ApoE* and *Dmd44*, where once again all the simulated R_M values are less than the actual value. All but one of the estimated rates C_{HRM} are greater than the approximate average for the human genome of 1 cM/Mb (e.g. Bouffard *et al.*, 1997; Nagaraja *et al.*, 1997). When 95% credibility intervals are constructed for C_{HRM} at each locus, the intervals exclude C_{map} for 6 of 9 loci (*β-globin*, *Duffy*, *Lpl*, *ApoE*, *Dmd44* and *Zfx*). When the loci are considered together, the C_{map} values as a whole can be strongly rejected ($X = 71{\cdot}7$; χ^2, 9 d.f.; $P < 10^{-11}$). Our finding of low levels of intragenic LD is particularly surprising in light of the high levels of pairwise linkage disequilibria that have been detected over hundreds of kilobases in some populations (e.g. Taillon-Miller *et al.*, 2000; Wilson & Goldstein, 2000). We now discuss possible explanations for the systematic differences between C_{HRM} and C_{map}.

https://doi.org/10.1017/S0016672301004967

Table 3. *The effect of a symmetric island model on the minimum number of recombination events,* R_M

Locus	$n_{Africans}$ [a]	$n_{Non\text{-}Africans}$ [b]	$F_{ST} = 0{\cdot}15$ P [c]	$F_{ST} = 0{\cdot}30$ P
β-globin	103	246	0·033	0·025
Duffy	48	34	0·009	0·007
Lpl	48	94	$< 10^{-5}$	$< 10^{-5}$
Ace	10	12	0·116	0·072
ApoE	48	144	$< 10^{-5}$	$< 10^{-5}$
Dmd44	10	31	$< 10^{-5}$	$< 10^{-5}$
Pdha1	16	19	0·126	0·091
Xq13·3	23	47	0·220	0·190
Zfx	114	222	0·029	0·025

[a] $n_{Africans}$ is the number of African chromosomes sampled in each survey.
[b] $n_{Non\text{-}Africans}$ is the number of non-African chromosomes sampled in each survey.
[c] $P = \Pr(R_M \geqslant \text{obs. } R_M | C_{map})$. To estimate P, we ran 10^5 simulations of a symmetric island model, with migrations rates chosen to yield an average F_{ST} value of 0·15 and 0·30 (cf. Takahata, 1983). P is the proportion of runs where R_M was greater than or equal to the observed value (listed in Table 1). In all cases, P decreases with decreasing migration rates, i.e. for a higher F_{ST} value.

(i) *Departures from demographic assumptions*

The discrepancy between estimates of C could be due to a number of factors, including error in our estimates of N_e or r, variation in recombination or mutation rates, multiple mutations at the same nucleotide site, and gene conversion. In addition, the demographic model underlying the calculation of C_{HRM} may be wrong in ways that lead to a substantial overestimate of C. Since the human population is now over 6 billion, an obvious candidate is a change in effective population size over time.

Evidence for an old onset to population growth (i.e. 50–100 Kya) is equivocal, with mtDNA, Y chromosome and most microsatellite data supporting it (e.g. Rogers & Harpending, 1992; Reich & Goldstein, 1998; Kimmel *et al.*, 1998; Gonser *et al.*, 2000; Thomson *et al.*, 2000) while many nuclear sequence studies do not (Hey, 1997; Harding *et al.*, 1997; Przeworski *et al.*, 2000; Wall & Przeworski, 2000). Table 2 presents estimates of C for a model of constant population size ($N_e = 10^4$) followed by 10-fold growth over 50 Kya. The discrepancy between C_{HRM} and C_{map} values is smaller, and 95% credibility intervals exclude C_{map} only for *β-globin* and *Dmd44*. For *Lpl*, C_{HRM} cannot be calculated because the estimated likelihood of the data is 0 under this model of exponential growth. In other words, for *any* recombination rate, the observed data is extremely likely. This may be due to an underestimate of the number of distinct haplotypes, as the phase of singletons and doubletons was not established. However, it is probably due at least in part to the inherent unlikeliness of the model of growth considered. In fact, at all 9 loci the likelihoods of C_{HRM} under recent population growth are less than the corresponding likelihoods under a constant population size model. By multiplying likelihoods across loci, we find ΠL (C_{HRM}) is more than 2×10^9 times larger under a constant size model than under our model of recent growth. If we were to fix the start time of exponential growth and consider the growth rate as a freely varying parameter, then 10-fold (or more) growth can be rejected at the 5% level for 5 of 9 loci. In the fourth column of Table 2, we list results for 100-fold growth instead of 10-fold; the qualitative conclusions are the same. In summary, even though recent exponential growth generally lowers the point estimates of C_{HRM}, the model of growth itself is extremely unlikely. Other aspects of human nuclear sequence data (e.g. the frequency spectrum of segregating mutations) also suggest that a simple model of growth is not appropriate (Wall & Przeworski, 2000).

A realistic demographic model is likely to be much more complex than models considered here (Wall & Przeworski, 2000). For example, population structure is well documented in humans (Cavalli-Sforza *et al.*, 1994). It should be noted, however, that population structure makes it even less likely to observe unusually high values of C_{HRM} relative to C_{map}. Subdivision decreases the effective rate of recombination since haplotypes in different subpopulations will not have a chance to recombine as often as under panmixia. This was verified by simulations of a symmetric finite-island model under a range of parameter combinations for the estimators of Hudson (1987), Griffiths & Marjoram (1996), Hey & Wakeley (1997) and Wall (2000) (results not shown). As an illustration, we run simulations under a two-island model (meant to correspond to African and non-African populations) for the 9 data sets in Table 1, with migration rates chosen to yield average F_{ST} values of 0·15 or 0·30 (cf. Takahata, 1983). Table 3 shows that in all cases, the probability of observing the actual value of R_M or

https://doi.org/10.1017/S0016672301004967

greater given C_{map} decreases, making the discrepancy between C_{HRM} and C_{map} more difficult to explain.

(ii) *Errors in parameter estimation*

A second explanation for the discrepancy between C_{HRM} and C_{map} is that N_e is underestimated. Estimates of N_e assume the standard neutral model, and are based on observed levels of diversity and an estimate of the per generation mutation rate (u). The rate u can be estimated directly, from the observed rate of spontaneous mutations, or indirectly, from divergence data, assuming a time (in generations) to the common ancestor of humans and a closely related species. The two methods yield similar estimates: when the first approach is applied to haemophilia B, $\hat{u} = 2{\cdot}14 \times 10^{-8}$ (Giannelli *et al*., 1999). Based on divergence between humans and chimpanzees at 18 pseudogenes, $\hat{u} = 2{\cdot}5 \times 10^{-8}$ (Nachman & Crowell, 2000*b*). The nucleotide diversity at 4-fold degenerate sites, which are thought to be evolving neutrally, is 0·11% (Li & Sadler, 1991; Cargill *et al*., 1999). Equating this level of diversity with $4N_e u$ for the two estimates of u, we obtain N_e estimates of $1{\cdot}1 \times 10^4$ and $1{\cdot}31 \times 10^4$. If these values are used instead of 10^4, C_{map} only increases by 10–31%. To assess whether errors in estimating μ or N_e could explain our observation, we redo our analyses with double the value of C_{map}. Seven of 9 loci have $C_{HRM} > 2C_{map}$, and $2C_{map}$ is excluded from the 95% credibility interval for 4 of 9 loci. As a whole, taking $2C_{map}$ as the recombination parameter at each locus can still be strongly rejected ($X = 50{\cdot}9$; χ^2, 9 d.f.; $P < 10^{-7}$).

A third possibility is that our estimates of the recombination rate per meiosis are in error. In this analysis, r is interpolated from the scale of megabases to one of kilobases. Different assumptions about the extent and scale on which rates vary across the genome lead to different choices for estimates of r. For example, if rates do not vary much and if genetic maps have large sampling errors, local estimates of r might best be obtained by averaging across many markers. In contrast, if they vary greatly, averaging over large distances might be uninformative locally. In Table 1, we list alternative estimates of r available for the loci considered; although far from exhaustive, these were obtained using a variety of maps and methods. Different methods yield rates that vary by as much as an order of magnitude. However, even if we take the highest estimate for each locus, C_{HRM} is greater than C_{map} at all 9 loci. This suggests that the pattern we observe is not the result of a particular choice of r estimate.

This said, the estimates listed in Table 1 are averages over large distances. Finer-scale mapping has yielded higher estimates for both the *β-globin* 5′ region and *Dmd44* (see below). Whatever method is used, estimates of r can only be extrapolated to different scales if variation in rates across the genome is negligible. Yet there is evidence of substantial variation in recombination rates across the genome at all scales (e.g. Fullerton *et al*., 1994; Dunham *et al*., 1999; Lien *et al*., 2000). In yeast and in maize, it has been proposed that some hotspots for recombination may be associated with promoter regions (cf. Atcheson & Easton Esposito, 1993; Lichten & Goldman, 1995; Dooner & Martinez-Ferez, 1997; Nicolas, 1998). If so, transcribed regions may experience higher rates of recombination than suggested by larger scale averages. (In Table 1, all but *Xq13·3* are transcribed.) Interestingly, *β-globin*, *Dmd44* and, more tentatively, *Lpl* are thought to contain a recombination hotspot (Oudet *et al*., 1992; Fullerton *et al*., 1994; Templeton *et al*., 2000), although only for *β-globin* and *Dmd44* is there evidence independent of polymorphism data. Our pattern remains even if we exclude *β-globin* and *Dmd44* from our analyses. For over half of the remaining loci, C_{map} falls outside the 95% credibility interval. When the information from the 7 remaining loci is combined, the C_{map} values can be excluded with high confidence ($X = 42{\cdot}2$; χ^2, 7 d.f.; $P < 10^{-6}$). This is still the case if all of the C_{map} values are doubled ($X = 26{\cdot}8$; χ^2, 7 d.f.; $P < 5 \times 10^{-4}$). Thus, our observation appears to be general, rather than the result of one or two fluke hotspots.

It is also interesting to note that the opposite pattern is found in *Drosophila*, where a sequence-based estimate of C (Hudson, 1987) is systematically *below* the laboratory-based estimate of the population rate of crossing-over per physical distance (Andolfatto & Przeworski, 2000). Thus, if genes are hotspots for recombination, they may be so in some species but not others.

(iii) *Other departures from model assumptions*

The simulations assume that all mutations occur at a previously unmutated site while in actual data there may be multiple hits to the same site. Violations of the infinite-sites model might lead to spurious inferences of recombination (cf. Templeton *et al*., 2000). For example, transitions away from CpG base pairs are thought to be 10–13 times more likely than other nucleotide base substitutions (Anagnostopoulos *et al*., 1999; Giannelli *et al*., 1999; Nachman & Crowell, 2000*b*; Templeton *et al*., 2000), and transversions away from CpG sites probably have elevated rates as well (Nachman & Crowell, 2000*b*). The fraction of polymorphisms that occur towards or away from CpG sites varies greatly across samples, from 0 at *Zfx* to 0·46 in *Pdha1*. What matters regarding inferences about recombination is not the number of such sites *per se*, but the expected number of multiple hits. Under the standard neutral model, this number can be

approximated as follows: Consider *Lpl*, with a heterozygosity level of 0·166%, excluding indels. (This estimate is conservative for our purposes, since we do not exclude CpG sites.) We find 20 polymorphisms at CpG sites (Templeton *et al*. (2000) report 19). If both transitions and transversions occur at 13 times the standard rate, an estimate of the population mutation rate, θ, away from CpG sites is roughly 2% per site (i.e. $0{\cdot}02 \approx 1{\cdot}66 \times 10^{-3} \times 13$). If we (conservatively) assume that all mutations occur away from CpGs, then the probability of a multiple hit is $\Pr(k \geqslant 2 | k \geqslant 1, \theta = 0{\cdot}02) \approx 0{\cdot}056$ per site. Thus, the expected number of multiple hits is roughly 1. Even if the multiple hit resulted in the spurious inference of two recombination events, the true R_M would be at least 20; the $\Pr(R_M \geqslant 20 | C = C_{map}) < 10^{-6}$. Thus, the expected number of multiple hits in these data sets seems sufficiently small to be of minor concern, unless there is a large degree of variability in mutation rates between different CpG sites. Even if we are extremely conservative and throw out all mutations at CpG sites, there is no noticeable trend on C_{HRM} values: four of them decrease, three of them increase, and the remaining two stay the same. Excluding all polymorphisms that occur at CpG sites, the 95% credibility intervals excluded C_{map} for 5 of 9 loci; they are generally broader since smaller data sets contain less information.

Another important factor in shaping intralocus patterns of linkage disequilibrium is gene conversion. Large-scale estimates of the recombination rate r do not include the effects of gene conversion; yet, on the scale of a gene, there may be an important contribution of gene conversion to the overall rate of genetic exchange (Andolfatto & Nordborg, 1998; Langley *et al*., 2000). We test this by implementing coalescent simulations with gene conversion (see Section 2). Table 4 lists the new estimates of C_{HRM} for the specific model considered. It is not surprising that with a model with more recombination (in the form of gene conversion), the differences between C_{map} and C_{HRM} decrease. Unlike the exponential growth simulations described earlier, L (C_{HRM}) generally increases when gene conversion is included; ΠL (C_{HRM}) increases 44-fold with the inclusion of gene conversion. In other words, the data (independent of C_{map}) indicate that a model of crossing-over and gene conversion actually fits the data better (i.e. is more likely) than a model of only crossing over. Both crossing-over and gene conversion tend to increase H and R_M; however, gene conversion generally increases R_M much more than H (results not shown). In human data we observe relatively large R_M values and relatively small H values, which is support for a model of gene conversion. However, our model of gene conversion is not a sufficient explanation for our data: the 95% credibility intervals still do not include C_{map} for 5 of 9 loci (*Lpl*, *β-globin*, *ApoE*, *Dmd44* and *Zfx*). The C_{map} values taken together can still be rejected ($X = 57{\cdot}0$; χ^2, 9 d.f.; $P < 10^{-8}$), even if *Dmd44* and *β-globin* are excluded ($X = 29{\cdot}1$; χ^2, 7 d.f.; $P < 2 \times 10^{-4}$). This is the case even if the ratio of conversion to crossover events is doubled (results not shown). We caution that very little is known about gene conversion in mammals, so it is unclear whether another model (e.g. one with a higher relative rate of gene conversion events), with the addition of hotspots at one or more loci, would be a sufficient explanation of our pattern.

Table 4. *Estimates of the population recombination rate under a model of crossing-over and gene conversion*

Locus	C_{HRM} (with gene conversion)[a]	$P^{a,b}$
β-globin	12·0	0·217
Duffy	2·3	0·061
Lpl	70·0	$< 10^{-5}$
Ace	9·2	0·519
ApoE	12·0	0·001
Dmd44	31·0	$< 10^{-5}$
Pdha1	2·9	0·490
Xq13·3	1·2	0·312
Zfx	1·6	0·062

[a] C_{HRM} and C_{map} refer to the population crossing-over rates (see text).
[b] $P = \Pr(R_M \geqslant \text{obs. } R_M | C_{map})$.

Estimates of C will only be useful if the model on which they depend is adequate. Yet, intragenic patterns of linkage disequilibrium appear to be inconsistent with both a constant population size model and a model of long-term exponential growth (given observed levels of diversity and divergence). A more appropriate demographic model should lead to some agreement between these and pedigree-based estimates of C. In principle, patterns of LD could be used to infer what a better model of human demography might be. However, this inference relies on a knowledge of local rates of recombination. If gene conversion and recombination hotspots are frequent within transcribed regions, pedigree-based estimates of the rate of crossing-over r may not be very informative locally, and intragenic patterns of LD may not help one infer the correct demographic model. Conversely, if independent evidence (such as the frequency spectrum of segregating mutations) suggests a model for the evolutionary past of humans, LD-based estimates of C will be quite informative about the local recombinational environment.

Unpublished data were kindly provided by Malia Fullerton, Martha Hamblin, Jody Hey, Gavin Huttley, Michael Nachman and Bret Payseur. We thank Peter Andolfatto, Dick Hudson, Trudy Mackay, Jonathan Pritchard and two anonymous reviewers for helpful comments. Most of this work was done while M. P. was supported by the University

of Chicago and J.D.W. by NIH grant #5 R01 HG10847. Both authors are currently supported by NSF postdoctoral grants in Bioinformatics.

References

Anagnostopoulos, T., Green, P. M., Rowley, G., Lewis, C. M. & Giannelli, F. (1999). DNA variation in a 5-Mb region of the X chromosome and estimates of sex-specific/type specific mutation rates. *American Journal of Human Genetics* **64**, 508–517.

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

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

Atcheson, C. L. & Easton Esposito, R. (1993). Meiotic recombination in yeast. *Current Opinion in Genetics and Development* **3**, 736–744.

Bouffard, G. G., Idol, J. R., Braden, V. V., Iyer, L. M., Cunningham, A. F., Weintraub, L. A., Touchman, J. W., Mohr-Tidwell, R. M., Peluso, D. C., Fulton, R. S., Ueltzen, M. S., Weissenbach, J., Magness, C. L. & Green, E. D. (1997). A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Research* **7**, 673–692.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalayanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q. & Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* **22**, 231–238.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press.

Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution* **7**, 111–122.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. & Sing, C. F. (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *American Journal of Human Genetics* **63**, 595–612.

Collins, A., Frezal, J., Teague, J. & Morton, N. E. (1996). A metric map of humans: 23500 loci in 850 bands. *Proceedings of the National Academy of Sciences of the USA* **93**, 14771–14775.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J. & Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature* **380**, 152–154.

Dooner, H. K & Martinez-Ferez, I. M. (1997). Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. *Plant Cell* **9**, 1633–1646.

Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K. N., Beasley, O., Bird, C. P., Blakey, S., Bridgeman, A. M., Buck, D., Burgess, J., Burrill, W. D. & O'Brien, K. P., *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402**, 489–495.

Fullerton, S. M., Harding, R. M., Boyce, A. J. & Clegg, J. B. (1994). Molecular and population genetic analysis of allelic sequence diversity at the human *β-globin* locus. *Proceedings of the National Academy of Sciences of the USA* **91**, 1805–1809.

Fullerton, S. M., Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Stengard, J. H., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. & Sing, C. F. (2000). Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *American Journal of Human Genetics* **67**, 881–900.

Giannelli, F., Anagnostopoulos, T. & Green, P. M. (1999). Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from hemophilia B. *American Journal of Human Genetics* **65**, 1580–1587.

Gonser, R., Donnelly, P., Nicholson, G. & Di Rienzo, A. (2000). Microsatellite mutations and inferences about human demography. *Genetics* **154**, 1793–1807.

Griffiths, R. C. & Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**, 479–502.

Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., Muselet, D., Prud'Homme, J. F., Dib, C., Auffray, C., Morissette, J., Weissenbach, J. & Goodfellow, P. N. (1996). A radiation hybrid map of the human genome. *Human Molecular Genetics* **5**, 339–346.

Hamblin, M. T. & Di Rienzo, A. (2000). Detecting the signature of natural selection in humans: evidence from the Duffy locus. *American Journal of Human Genetics* **66**, 1669–1679.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., Cox, M. J., Schneider, J. A., Moulin, D. S & Clegg, J. B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *American Journal of Human Genetics* **60**, 772–789.

Harris, E. & Hey, J. (1999). X chromosome evidence for ancient human histories. *Proceedings of the National Academy of Sciences of the USA* **96**, 3320–3324.

Hey, J. (1997). Mitochondrial and nuclear genes present conflicting portraits of human origins. *Molecular Biology and Evolution* **14**, 166–72.

Hey, J. & Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846.

Hilliker, A. J., Harauz, G., Reaume, A. G., Gray, M. & Clark, S. H. (1994). Meotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. *Genetics* **137**, 1019–1026.

Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research* **50**, 245–250.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **1**, 1–14.

Hudson, R. R. (1993). The how and why of generating gene genealogies. In *Mechanisms of Molecular Evolution* (ed. N. Takahata & A. G. Clark), pp. 23–36. Sunderland, MA: Sinauer Associates.

Hudson, R. R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J., Castle, A. B., Silva, J., Slonim, D. K., Baptista, R., Kruglyak, L. & Xu, S. H., *et al.* (1995). An STS-based map of the human genome. *Science* **270**, 1945–1954.

Huttley, G. M., Smith, M. W., Carrington, M. & O'Brien, S. J. (1999). A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722.

Jaruzelska, J., Zietkiewicz, E., Batzer, M., Cole, D. E., Moisan, J. P., Scozzari, R., Tavare, S. & Labuda, D. (1999). Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. *Genetics* **152**, 1091–1101.

Kaessmann, H., Heissig, F., Von Haessler, A. & Paabo, S. (1999). DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nature Genetics* **22**, 78–81.

Kimmel, M., Chakraborty, R., King, J. P., Bamshad, M., Watkins, W. S. & Jorde, L. B. (1998). Signatures of population expansion in microsatellite repeat data. *Genetics* **148**, 1921–1930.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Applications* **13**, 235–248.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22**, 139–144.

Langley, C. H., Lazzaro, B. P., Phillips, W., Heikkinen, E. & Braverman, J. M. (2000). Linkage disequilibria and the site frequency spectra in the *su*(*s*) and *su*(w^a) regions of the *Drosophila melanogaster* X chromosome. *Genetics* **156**, 1837–1852.

Lichten, M. & Goldman, A. S. H. (1995). Meiotic recombination hotspots. *Annual Review of Genetics* **29**, 423–444.

Li, W. H. & Sadler, L. (1991). Low nucleotide diversity in man. *Genetics* **129**, 513–523.

Lien, S., Szyda, J., Schechinger, B., Rappold, G. & Arnheim, N. (2000). Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *American Journal of Human Genetics* **66**, 557–566.

Long, A. D. & Langley, C. H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* **9**, 720–731.

Marjoram, P. & Donnelly, P. (1994). Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* **136**, 673–683.

Nachman, M. W. & Crowell, S. (2000*a*). Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. *Genetics* **155**, 1855–1864.

Nachman, M. W. & Crowell, S. (2000*b*). Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304.

Nachman, M. W., Bauer, V. L., Crowell, S. L. & Aquadro, C. F. (1998). DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141.

Nagaraja, R., MacMillan, S., Jones, C., Masisi, M., Pengue, G., Porta, G., Miao, S., Casamassimi, A., D'Urso, M., Brownstein, B. & Schlessinger, D. (1997). X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Research* **7**, 210–222.

Nicolas, A. (1998). Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. *Proceedings of the National Academy of Sciences of the USA* **95**, 87–89.

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

Orr-Weaver, T. L. & Szostak, J. W. (1985). Fungal recombination. *Microbiological Reviews* **49**, 33–58

Oudet, C., Hanauer, A., Clemens, P., Caskey, T. & Mandel, J. L. (1992). Two hot spots of recombination in the *Dmd* gene correlate with the deletion prone regions. *Human Molecular Genetics* **1**, 599–603.

Payseur, B. A. & Nachman, M. W. (2000). Microsatellite variation and recombination rate in the human genome. *Genetics* **156**, 1285–1298.

Petes, T. D., Malone, R. E. & Symington, L. S. (1991). Recombination in yeast. In *The Molecular and Cellular Biology of the Yeast* Saccharomyces*: Genome Dynamics, Protein Synthesis, and Energetics*, vol. 1 (ed. J. R. Broach, J. R. Pringle & E. W. Jones), pp. 407–521. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Przeworski, M., Hudson, R. R. & Di Rienzo, A. (2000). Adjusting the focus on human variation. *Trends in Genetics* **16**, 296–302.

Reich, D. & Goldstein, D. (1998). Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the USA* **95**, 8119–8123.

Rieder, M. J., Taylor, S. L., Clark, A. G. & Nickerson, D. A. (1999). Sequence variation in the human angiotensin converting enzyme. *Nature Genetics* **22**, 59–62.

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

Rogers, A. R. & Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**, 552–569.

Taillon-Miller, P., Bauer-Sardiña, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P. & Kwok, P.-Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human *Xq25* and *Xq28*. *Nature Genetics* **25**, 324–328.

Takahata, N. (1983). Gene identity and genetic differentiation of populations in the finite island model. *Genetics* **104**, 497–512.

Takahata, N. (1993). Allelic genealogy and human evolution. *Molecular Biology and Evolution* **10**, 2–22.

Templeton, A. R., Clark, A. G., Weiss, K. M., Nickerson, D. A., Boerwinkle, E. & Sing, C. F. (2000). Recombinational and mutational hotspots within the human Lipoprotein Lipase gene. *American Journal of Human Genetics* **66**, 69–83.

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. & Feldman, M. W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proceedings of the National Academy of Sciences of the USA* **97**, 7360–7365.

Wall, J. D. (2000). A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution* **17**, 156–163.

Wall, J. D. & Przeworski, M. (2000). When did the human population size start increasing? *Genetics* **155**, 1865–1874.

Wilson, J. F. & Goldstein, D. B. (2000). Consistent long-range linkage disequilibrium generated by admixture in a Bantu–Semitic hybrid population. *American Journal of Human Genetics* **67**, 926–935.

Wiuf, C. & Hein, J. (2000). The coalescent with gene conversion. *Genetics* **155**, 451–462.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.