Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization

Magnus Nordborg

Department of Genetics, Lund University, S-223 62 Lund, Sweden

Manuscript received March 26, 1999
Accepted for publication October 11, 1999

ABSTRACT

It is shown that partial self-fertilization can be introduced into neutral population genetic models with recombination as a simple change in the scaling of the parameters. This means that statistical and computational methods that have been developed under the assumption of random mating can be used without modification, provided the appropriate parameter changes are made. An important prediction is that all forms of linkage disequilibrium will be more extensive in selfing species. The implications of this are discussed.

THE tremendous advances in methods of polymorphism detection, especially directly at the sequence level, are in the process of rejuvenating population genetics (Chakravarti 1999). To fully take advantage of the flood of polymorphism data, it is clear that new models and methods of analysis are needed. In particular, there is a need for theory that incorporates recombination, so that phenomena such as linkage disequilibrium can be investigated. Although multilocus models have a long history, most of this work concerns small numbers of loci, whereas sequences, in which each base can be seen as a locus, consist of very large numbers of loci. Indeed it can be argued that the current focus on linkage disequilibrium in the sense of a measure of correlation between pairs of loci is a relic of classical two-locus models.

Furthermore, to be useful in analyzing data, theory must be formulated in an appropriate statistical framework. Many well-known results in population genetics are expectations over evolutionary realizations, whereas data typically reflect a single such realization (Donnelly 1996). Models based on gene genealogies, in particular the neutral coalescent, have played an important role in developing theory suitable for analyzing data (Hudson 1990; Donnelly and Tavaré 1995), and, although the standard coalescent does not include recombination, it can readily be generalized to do so (Hudson 1983).

Classical frequency-based models tend to be limited to small numbers of loci to keep the dimensionality of the model manageable. For the same reason, random mating, in the sense of random union of gametes, is usually assumed. This makes it possible to describe the model in terms of gamete frequencies rather than genotype frequencies, thus avoiding the complications of diploidy (Ewens 1979). The standard coalescent has usually also been studied in what is essentially a haploid setting, but it has recently been shown that diploidy and nonrandom mating can be incorporated with surprising ease (Nordborg and Donnelly 1997; Möhle 1998). In this article, I show that a similar, strikingly simple result holds for the more general class of genealogical models that allow recombination, and I discuss the implications of this for the analysis and utilization of polymorphism data.

Address for correspondence: Department of Genetics, Lund University, Sölvegatan 29, S-223 62 Lund, Sweden.
E-mail: magnus.nordborg@gen.lu.se

RESULTS

I first review how the standard coalescent has been extended to allow recombination and selfing separately, and then show how both may be included simultaneously. The discussion of recombination is based on the "ancestral recombination graph" described by Griffiths and Marjoram (1997), but the results apply equally well to the alternative formulations of Hudson (1983) or Wiuf and Hein (1999). The coalescent with selfing is described in Nordborg and Donnelly (1997), and a more formal treatment is given in Möhle (1998).

Recombination: The standard coalescent is based on the simple argument that, looking backward in time, each gene copy can be seen as "picking" its parent randomly from among the copies that were present in the preceding generation (note that this requires neutrality). Thus lineages can be traced back in time: whenever two lineages pick the same parent, a coalescence event occurs, and the number of lineages is reduced by one, until eventually a single lineage, the most recent common ancestor (MRCA), is found. Recombination is readily introduced into this framework: looking backward in time, a recombination event in a given lineage simply means that that lineage splits into two separate lineages, one being ancestral for the segment to the left

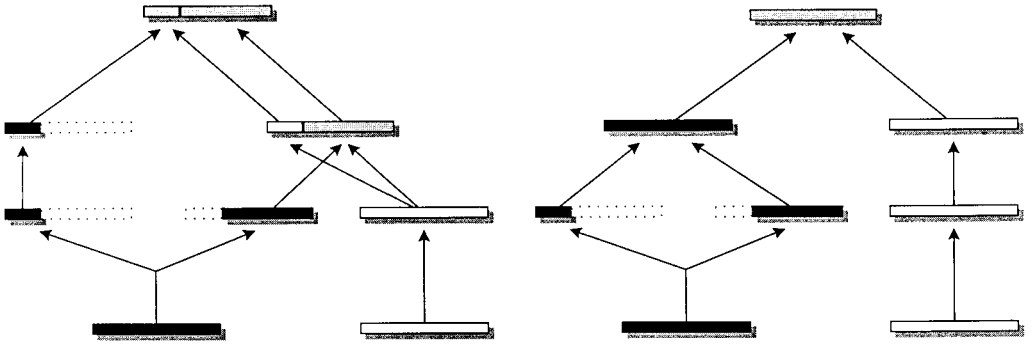

Figure 1.—Illustration of the coalescent with recombination. Two different genealogical graphs for two segments (labeled black and white) are depicted. Note that the colors are simply labels denoting ancestry; thus gray segments are ancestral to both black and white, whereas "dotted" segments are ancestral to neither. In both graphs, the first (most recent) event is a recombination event that creates two separate lineages for the black segment, and in the graph on the left, this induces different genealogical trees on either side of the breakpoint. However, in the graph on the right, the recombination event ends up being entirely undetectable, because the two products coalesce back together before either coalesces with another ancestral lineage.

of the breakpoint, the other being ancestral for the segment to the right of the breakpoint (Figure 1).

More formally, consider n chromosomal segments of length r M, sampled from a large diploid of size N monoecious (hermaphroditic) individuals. The population is assumed to evolve in discrete time according to the neutral Wright-Fisher model, but the standard diffusion (continuous) time scaling is employed, *i.e.*, time is measured in units of $2N$ generations, while $N \to \infty$. When this is done, each pair of lineages can be shown to coalesce at rate 1, so that, in the absence of recombination, the total number of ancestral lineages decreases back in time as a pure death process with rate $k(k-1)/2$, where $k = n, n-1, \ldots, 1$. However, with recombination, the number of lineages may also increase (*cf.* Figure 1). Letting R be the scaled recombination rate, obtained by holding $R = 4Nr$ fixed as $N \to \infty$, each ancestral lineage undergoes recombination at rate $R/2$, so that the total number of lineages increases at rate $kR/2$. Because the coalescence (death) rate is quadratic and the recombination (birth) rate is linear, the process is guaranteed to coalesce to a single lineage, the "ultimate" MRCA, in finite time. However, the genealogy at each point is described by an embedded standard coalescent, and every single point may well have coalesced to its MRCA before the ultimate MRCA is reached.

Selfing: The simplest way to think of a diploid population of size N is as a haploid population of size $2N$ subdivided into N islands of size 2. If time is rescaled in units of $2N$ generations as described above, then, looking backward in time, each pair of lineages "coalesces" into the same island at rate 2. Whenever this happens, there are two possibilities: either the two lineages pick the same of the two available parents, or they pick different ones. The former event, which occurs with probability 1/2, results in a real coalescence, whereas the latter event, which also occurs with probability 1/2, simply results in the two lineages temporarily occupying the same island. If the population is outcrossing (random mating), two lineages currently on the same island pick their islands in the preceding generation independently of each other, just like lineages occupying different islands. Real coalescences thus occur at rate $2 \times 1/2 = 1$ for each pair, as in the standard coalescent.

However, if the population is partially selfing, so that a fraction s of all zygotes are produced through self-fertilization and a fraction $1-s$ through outcrossing, then two lineages currently on the same island (*i.e.*, in the same individual) do not pick islands in the preceding generation independently of each other. Instead, with probability s, they automatically pick the same island, and with probability $1-s$, they pick islands independently of each other. If they pick the same island, the probability is again 1/2 that they pick the same parent and coalesce and 1/2 that they pick different ones and remain distinct. Continuing back in time, it is easy to see that two lineages starting on the same island will, in a small number of generations, either coalesce or end up picking islands independently of each other. The former event occurs with probability

$$\frac{s/2}{s/2 + 1 - s} = \frac{s}{2 - s} \equiv F$$

and the latter with probability $1 - F$. Thus, each time a pair of lineages pick the same island, the total probability that this event results in a coalescence is

$$\frac{1}{2} \times 1 + \frac{1}{2} \times F = \frac{1 + F}{2}.$$

On the diffusion time scale, the rate at which each pair picks the same island is 2, as we have seen, and the amount of time spent by two distinct lineages on the same island is negligible, so the total rate of coalescence for each pair becomes simply $1 + F$.

Thus, the coalescent with partial selfing looks like the standard coalescent, but with a rate of coalescence that is $1 + F$ times faster. Alternatively, if time is rescaled in

units of $2N/(1 + F)$ generations (the so-called "effective population size"; *cf.* Pollak 1987), the process looks exactly like the standard coalescent. The latter formulation has the important consequence that all statistical and computational methods that have been developed for the standard coalescent can be used without modification by simply rescaling the mutation parameter. For example, a computer simulation written to generate random samples with a certain mutation rate Θ from an outcrossing population can also generate equivalent samples from a partially selfing one by simply replacing Θ with an "effective mutation rate" $\Theta_s \equiv \Theta/(1 + F)$.

Note that, with selfing, consideration needs to be given to whether both copies are sampled from each diploid individual or not, because, as we have seen, pairs within the same individual instantaneously coalesce with probability F. This corresponds to the well-known increase in the frequency of homozygotes in partial selfers and is easily taken into account (Nordborg and Donnelly 1997).

Recombination and selfing: Given the two separate processes, the behavior of the coalescent with *both* recombination and selfing can be understood with just a single additional insight. As illustrated in Figure 1, a recombination event has absolutely no effect on the genealogical graph if the two pieces containing the break point coalesce with each other before either of them coalesces with another lineage (Hudson and Kaplan 1985). It is evident from Figure 1 that this is not an unlikely event for small samples; in fact, the probability that a given break is eventually "healed" by coalescence in this manner can be shown to be

$$p(n) = \frac{2(n-1)}{3n\sum_{k=1}^{n-1} k^{-1}} \sim \frac{2}{3}\frac{1}{\gamma + \log(n)} \quad \text{as } n \to \infty$$

in the standard model (see the appendix for proof). However, much more striking is the fact that, with selfing, a fraction F of all recombination events are healed *instantaneously*. This follows directly from the argument presented above: the two recombinants obviously occupy the same individual and thus coalesce instantaneously with probability F. In other words, the ancestral recombination graph with partial selfing looks like the outcrossing version but with a rate of coalescence that is $1 + F$ times faster and a rate of recombination that is $1 - F$ times slower (as noted above, some additional considerations apply if both copies are sampled from the same individual). Alternatively, if time is rescaled in units of $2N/(1 + F)$ generations, the process looks like the outcrossing version but with a recombination rate that is $(1 - F)/(1 + F) = 1 - s$ times slower. There is thus no longer a single "effective population size" that recaptures the original model, but it is possible to accomplish this by also introducing an "effective recombination rate." Again, an important consequence of the second formulation is that methods and results derived for outcrossing can be reused if Θ is replaced by Θ_s and R by $R_s \equiv R(1 - s)$.

These results correspond to the classical "forward" intuition that recombination will be less effective in eliminating linkage disequilibrium in selfing organisms because more individuals are homozygotes. Indeed, Golding and Strobeck (1980) showed that the results hold for the expectation of various squared pairwise linkage disequilibria in a two-locus infinite-alleles model. The present results thus confirm and considerably extend their results. Related arguments have also been used to derive an "effective recombination rate" in the context of strain evolution in malaria parasites (Dye and Williams 1997).

Because of the confusion about "effective population size" in population genetics, it is worth emphasizing the generality of the results. Introducing the "effective" rates of mutations and recombination, Θ_s and R_s, respectively, does not merely cause the expectation of some functional to behave as in the outcrossing model; rather, it causes the full coalescent process to behave identically (for a more rigorous formulation, see Möhle 1998). By the close relationship between coalescent and diffusion processes, the results should also hold for relevant diffusion models (*e.g.*, Ethier and Griffiths 1991).

SIMULATED EXAMPLES

I use standard Monte Carlo simulation methods to illustrate the implications of the results. The principle behind the simulation is simple: generate a random graph according to the chosen recombination model, then "drop" mutations onto the graph according to the chosen mutation model at rate $\Theta/2$ along the branches. I have assumed that both types of events occur at points picked uniformly from a real interval. This corresponds to the classical infinite-sites mutation model (Kimura 1969); by analogy, it is natural to refer to the recombination model used here as "infinite-sites." The assumptions may be reasonable for DNA sequences, because the probability of an event per generation per base pair is very low. However, it should be emphasized that the ideas developed in this article are quite general and allow a wide range of models (some mutation models are discussed in Donnelly and Tavaré 1995).

Figures 2 and 3 show summary statistics for two samples generated with the same per base pair mutation and recombination parameters but assuming either outcrossing (Figure 2) or 95% selfing (Figure 3). I have furthermore assumed that the per base pair rates of recombination and mutation, ρ and θ, respectively, are equal. This is of course not true generally, and will certainly vary greatly between genomic regions, but a one-to-one ratio nonetheless seems to be reasonable for both Drosophila (Hudson 1987) and humans (Clark

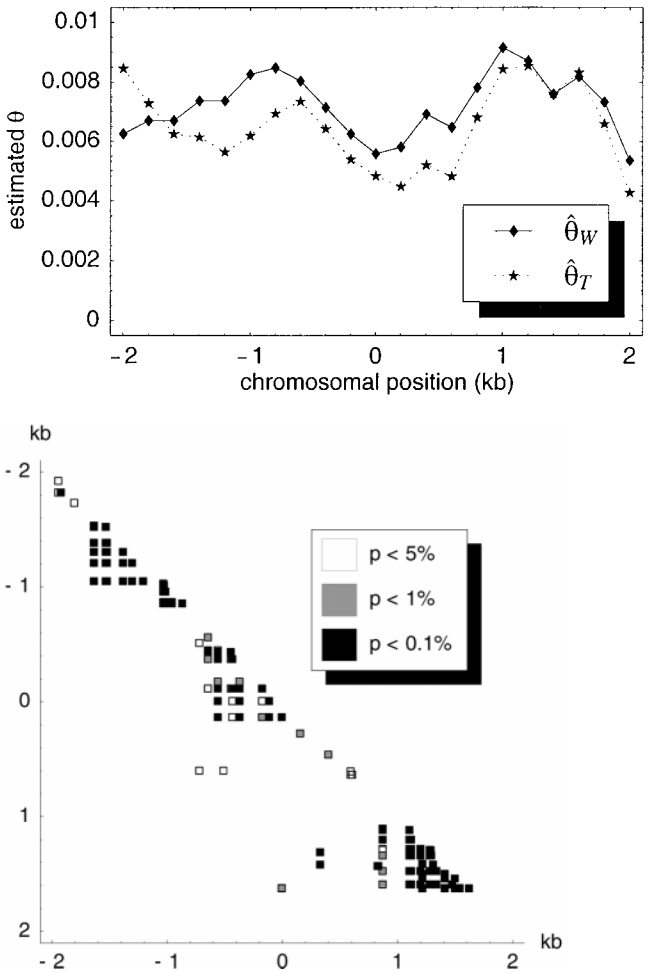

Figure 2.—A simulated sample of $n = 50$ sequences from an outcrossing ($s = 0$) population, summarized in two popular ways. The sample was generated as described in the text, with $\Theta = R = 40$. Under the assumption that $\theta = 0.01$, this corresponds to a sequence length of 4 kb. The top plot shows estimates of θ, obtained using moving averages with a window size of 1 kb, and two different estimators, $\hat{\theta}_W$ (Watterson 1975) and $\hat{\theta}_T$ (Tajima 1983). The bottom plot shows all pairwise linkage disequilibria that have probabilities <5% under Fisher's exact test with Bonferroni correction for multiple comparisons (Sokal and Rohlf 1981). Darker color corresponds to lower probability.

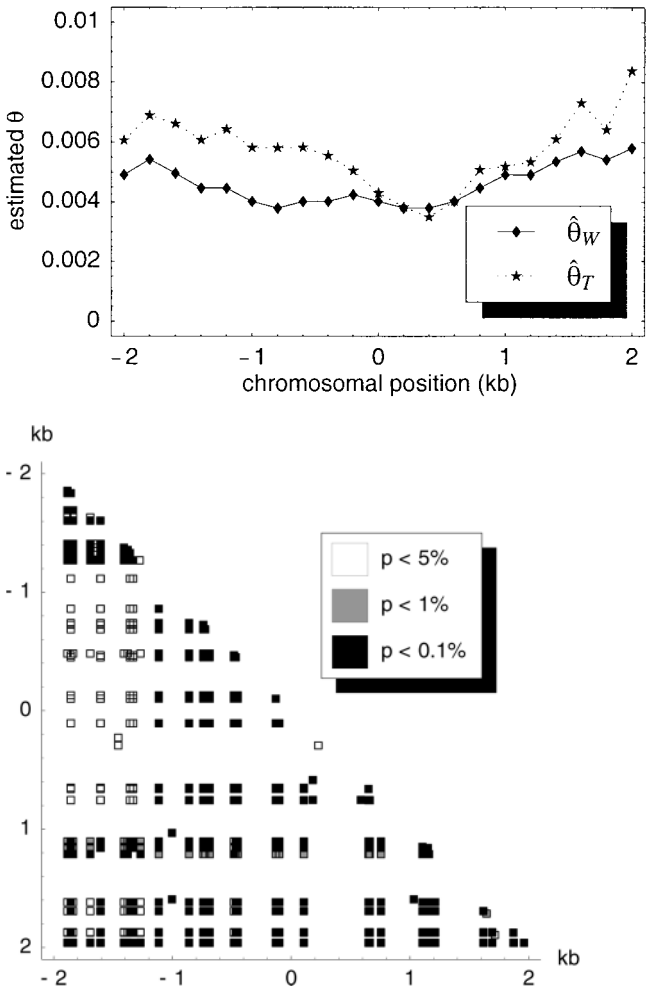

Figure 3.—A simulated sample of $n = 50$ sequences, from 50 different individuals, in a highly selfing ($s = 0.95$) population, summarized as described in the legend to Figure 2. The sample was generated using the same method as in Figure 2, assuming the same per base pair parameters, so that $\Theta_s = 40(1 - 0.95/2) = 21$ and $R_s = 40(1 - 0.95) = 2$.

et al. 1998). Because the model is formulated in terms of the total rate parameters Θ and R, it is also necessary to decide what these parameters mean in terms of actual physical distances. The number of base pairs corresponding to a certain total mutation rate Θ is of course Θ/θ. In the examples, I have assumed $\theta = 0.01$, which is approximately correct for Drosophila, but probably an order of magnitude too large for humans (Li 1997). However, the simulated samples could equally well be thought of as originating from organisms more like humans than like Drosophila if one were to multiply the scale on the axes representing physical distance in Figures 2 and 3 by a factor of 10 and divide the estimates of θ by another factor of 10.

DISCUSSION

Effects on the level of polymorphism: Consider first the estimates of θ, shown in the top plots of Figures 2 and 3. These estimates are commonly thought of as measures of the "level of polymorphism." The expectation of the estimators is $\theta = 0.01$ in Figure 2 and $\theta_s = \theta/(1 + F) \approx 0.005$ in Figure 3: an almost twofold difference. However, the estimates in this pair of samples do not differ nearly as much: indeed, for part of the sequence, θ is estimated to be higher in the selfer than in the outcrosser. It is important to realize that this is not unusual. Estimators of θ are well known to have extremely large variances, mostly due to the randomness of the underlying genealogy. With respect to this "historical" or "evolutionary" variance, a population sample from a single nonrecombining locus is essentially a sample of size one (Donnelly and Tavaré 1995; Donnelly 1996). Recombination improves the estimates, because

it breaks the sample up into different genealogical trees (the graphs underlying the data in Figures 2 and 3 consisted of 144 and 9 distinct trees, respectively), but these trees are still strongly correlated (Pluzhnikov and Donnelly 1996).

The predicted difference between comparable selfing and outcrossing species is thus impossible to detect without data from many loci and is likely to be often obscured by other demographic factors in any case. The fact that samples from selfers often seem to yield considerably smaller estimates of θ is probably best explained through founder effects and/or selection at linked loci (Cummings and Clegg 1998; Liu *et al.* 1999).

Effects on linkage disequilibrium: Consider next the correlations between the genealogical trees along the chromosome. The bottom plots of Figures 2 and 3 show the pattern of linkage disequilibrium between all pairs of polymorphic sites as a function of their probability under Fisher's exact test. While the relationship between the pairwise linkage disequilibrium coefficient and correlations in tree topologies is not understood, it is clear that both decrease with recombinational distance. The patterns of pairwise linkage disequilibrium differ strikingly between the outcrossing and selfing sample: in the former, strong linkage disequilibrium is essentially only in evidence within distances of <500 bp, and it completely vanishes over the entire 4-kb region, whereas in the latter, no decay with distance is evident within the 4-kb region. In contrast to the pattern discussed above, this difference between outcrossing and highly selfing samples is to be expected (M. Nordborg, unpublished simulation results). That this should be so is perhaps not surprising in light of the fact that whereas the effective mutation rate is reduced from 40 to 21 (a factor of almost 2), the effective recombination rate is reduced from 40 to 2 (a factor of 20). Available data from outcrossing (*e.g.*, Clark *et al.* 1998; Long *et al.* 1998) and highly selfing (reviewed in Cummings and Clegg 1998; Liu *et al.* 1999) organisms are in general agreement with this predicted difference in the decay of linkage disequilibrium.

Given that linkage disequilibrium seems to vanish over a scale of 500 bp in Figure 2 but shows little sign of vanishing over 4 kb in Figure 3, it is worth asking over what scale it would vanish in the latter case. A rough answer can be obtained by noting that, in Figure 2, 500 bp correspond to $R = 5$, and that, in Figure 3, $R_s = 5$ corresponds to $R = 100$, which in turn corresponds to 10 kb. Linkage disequilibrium would thus vanish over a scale of 10 kb for the parameters in Figure 3. This ignores the fact that the number of polymorphic sites would also change, which affects the power to detect linkage disequilibrium.

The relevance and number of trees: Another, quite illuminating, way of thinking of the effects of selfing on linkage disequilibrium is to consider the underlying graph as given and then interpret it for different rates

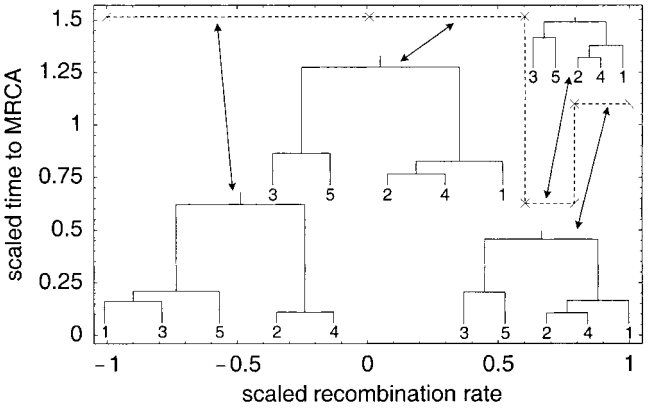

Figure 4.—A simulated recombination graph for a sample of five segments, with $R = 2$. The dotted line shows the total tree height for each chromosomal position. The crosses along this line indicate points where a recombination event induced a new genealogical tree. These embedded trees are shown next to their region and are drawn to scale with respect to each other. From left to right, the tree topology changes between the first and the second trees, whereas the last three trees have identical topologies but very different branch lengths. This particular realization was selected for clarity.

of selfing by rescaling the mutation rate. Figure 4 shows a particular simulated graph for a sample of $n = 5$ segments, with $R = 2$. The graph consists of four embedded trees. In general, the expected number of embedded trees will increase with R and n, but the distribution is not known: Griffiths and Marjoram (1997) show that the expected logarithm of the number of trees is asymptotic to $R(\log 2)(\log n)$ as $n \to \infty$; an expression for the expected number of recombination events that affects the graph can be found in Hudson and Kaplan (1985); and expressions for the expected numbers affecting the graph in different ways have also been obtained (C. Wiuf, personal communication). In any case, four trees is not an abnormal number for $n = 5$, $R = 2$.

Imagine first that the graph in Figure 4 comes from a species with the "Drosophila" parameters used in Figure 2. Then $R = 2$ corresponds to a physical distance of 200 bp, and mutations should be added with $\Theta = 2$. In other words, the expected number of segregating sites in the sample would only be approximately four, and there is obviously no hope of ever reconstructing the four embedded trees. In a very practical sense, the data have no tree structure.

If we instead assume the "human" parameters described above, then $R = 2$ would correspond to 2 kb; however, there would still only be an average of four segregating sites. Thus each tree would correspond to a larger chromosomal region, but there would still not be enough polymorphism to reconstruct them.

To increase the power to infer the underlying genealogical graph, the ratio between the mutation and recombination rates needs to be increased. This cannot be done by simply changing the population size, because

it appears in both numerator and denominator. It can be done by changing the mating system, however, because

$$\frac{\Theta_s}{R_s} = \frac{\Theta}{(1 - F)\, R}. \quad (1)$$

Imagine that the graph in Figure 4 comes from a selfing population with the parameters in Figure 3. Then the recombination rate of 2 used in generating Figure 4 would be the same as that used in generating Figure 3, and the corresponding region would therefore be 4 kb and $\Theta_s = 21$. The expected number of segregating sites would be 44, the data would look like those summarized in Figure 3, and there would be considerable information about the embedded trees.

The relevance of the recombination model: The example just presented calls into question whether the recombination model is appropriate for outcrossing organisms with large population size. Like most models of recombination in population genetics, the one used here assumes that recombination is a point event and that the probability of such an event between two points decreases linearly to zero as the distance between the points decreases to zero. Yet it is well known that recombination is mechanistically tied to gene conversion. Ignoring this is reasonable when modeling distances large enough for the effects of gene conversion to be negligible (Andolfatto and Nordborg 1998). However, the average gene conversion tract in Drosophila seems to be on the order of 300 bp, and the graph in Figure 4 has four different trees within 200 bp.

Similarly, it is worth questioning whether the assumption that the probability of recombination is independent of the actual sequence is reasonable for such small distances. Many types of recombination heterogeneity pose no problem from a mathematical point of view, but the lack of data to guide the modeling does.

Implications for linkage disequilibrium mapping: There is currently tremendous interest in using linkage disequilibrium to map genes. The results presented here might seem to indicate that this would not be possible in outcrossing organisms, because linkage disequilibrium is only expected between extremely tightly linked sites. Yet linkage disequilibrium has been used repeatedly to map human diseases. At least part of the explanation lies in the fact that the linkage disequilibrium plotted in Figures 2 and 3 is unconditional, between random polymorphisms, whereas the human disease cases have invariably utilized association between markers and a given site that is known to be rare, thus young, so that significant traces of its ancestral haplotype still persist. It remains to be seen whether linkage disequilibrium mapping will work more generally; the pattern of linkage disequilibrium surrounding older, more frequent mutations could well look more like the patterns above (Kruglyak 1999). The present results do suggest that linkage disequilibrium mapping may, in some circumstances, work much better in partially selfing species like Arabidopsis, rice, barley, or *Plasmodium falciparum* (Conway *et al.* 1999) than in outcrossing ones like humans or Drosophila.

Estimating the selfing rate: Milligan (1996) suggested that sequence data could be used to estimate the long-term mating system, essentially by comparing estimates of θ within and between individuals. Nordborg and Donnelly (1997) showed that this suggestion was incorrect; almost all the information utilized by the proposed estimators comes from the increased frequency of homozygotes, which reflects only the mating pattern of the last couple of generations. However, the model used ignored recombination, and it is obvious from comparing Figures 2 and 3 that there is information about the selfing rate in the pattern of linkage disequilibrium. This pattern does reflect the long-term mating system as linkage disequilibrium builds up and decays slowly. From Equation 1 we see that a rough estimate of the long-term mating system can be obtained by simply dividing an estimate $\hat{\Theta}$ by an estimate $\hat{R}$, assuming some external knowledge of the ratio between the per-base pair probabilities of mutation and recombination. Available data do suggest that the ratio $\hat{\Theta}/\hat{R}$ is at least an order of magnitude higher in highly selfing than in outcrossing species (Cummings and Clegg 1998). Better estimators could no doubt be found (following, *e.g.*, Griffiths and Marjoram 1996); however, it is worth considering whether any inference of this type would be robust to the effects of population structure, which is likely to be an important factor in most partially selfing organisms.

The power of coalescence arguments: Finally, the results of this article strikingly demonstrate the potential power of coalescence arguments. To study recombination in a classical diffusion setting, a minimum of three dimensions are needed even if random mating is assumed. With partial selfing, the minimum dimensionality is increased to nine. The results of Golding and Strobeck (1980) were obtained using a system of 16 identity coefficients and a considerable amount of algebra. Mutation models appropriate for modern data require infinite-dimensional, measure-valued diffusion processes (Ethier and Griffiths 1987). In contrast, by looking backward in time, a general result becomes obvious.

I thank Bengt Bengtsson, Deborah Charlesworth, Andy Clark, Peter Donnelly, Bob Griffiths, Carsten Wiuf, and two anonymous reviewers for comments on the manuscript. This work was supported by the Swedish Natural Sciences Research Council (NFR Grant B-AA/BU 12026) and by the Erik Philip-Sörensen Foundation.

LITERATURE CITED

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

Chakravarti, A., 1999 Population genetics—making sense out of sequence. Nat. Genet. **21:** 56–60 (suppl).
Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.
Conway, D. J., C. Roper, A. M. J. Oduola, D. E. Arnot, P. G. Kremsner *et al.*, 1999 High recombination rate in natural populations of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **96:** 4506–4511.
Cummings, M. P., and M. T. Clegg, 1998 Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (*Hordeum vulgare* ssp. *spontaneum*): an evaluation of the background selection hypothesis. Proc. Natl. Acad. Sci. USA **95:** 5637–5642.
Donnelly, P., 1996 Interpreting genetic variability: the effects of shared evolutionary history, pp. 25–50 in *Variation in the Human Genome*. Wiley, Chichester, United Kingdom.
Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.
Dye, C., and B. G. Williams, 1997 Multigenic drug resistance among inbred malaria parasites. Proc. R. Soc. Lond. Ser. B **264:** 61–67.
Ethier, S. N., and R. C. Griffiths, 1987 The infinitely many sites model as a measure-valued diffusion. Ann. Prob. **5:** 515–545.
Ethier, S. N., and R. C. Griffiths, 1991 The neutral two-locus model as a measure-valued diffusion. Adv. Appl. Prob. **22:** 773–786.
Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.
Golding, G. B., and C. Strobeck, 1980 Linkage disequilibrium in a finite population that is partially selfing. Genetics **94:** 777–789.
Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comp. Biol. **3:** 479–502.
Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavaré. Springer-Verlag, New York.
Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.
Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.
Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–43 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyma and J. Antonovics. Oxford University Press, Oxford.
Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.
Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.
Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.
Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
Liu, F., D. Charlesworth and M. Kreitman, 1999 The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus Leavenworthia. Genetics **151:** 343–357.
Long, A. D., R. F. Lyman, C. H. Langley and T. F. C. Mackay, 1998 Two sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. Genetics **149:** 999–1017.
Milligan, B. G., 1996 Estimating long-term mating systems using DNA sequences. Genetics **142:** 619–627.
Möhle, M., 1998 A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. Adv. Appl. Prob. **30:** 493–512.
Nordborg, M., and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146:** 1185–1195.
Pluzhnikov, A., and P. Donnelly, 1996 Optimal sequencing strategies for surveying molecular genetic diversity. Genetics **144:** 1247–1262.
Pollak, E., 1987 On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117:** 353–360.
Saunders, I. W., S. Tavaré and G. A. Watterson, 1984 On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. **16:** 471–491.
Sokal, R. R., and F. J. Rohlf, 1981 *Biometry*. W. H. Freeman, New York.
Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.
Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.
Wiuf, C., and P. Donnelly, 1999 Conditional genealogies and the age of a neutral mutant. Theor. Popul. Biol. **56:** 183–201.
Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. **55:** 248–259.

Communicating editor: A. G. Clark

APPENDIX

We seek the probability that a given recombination break is "healed" by coalescence, as depicted in the right-hand graph of Figure 1. Note first that the fate of the break depends only on the genealogy at the break point; other recombination events are irrelevant. Suppose the recombination event occurred while there were l lineages ancestral to this point. Then the break is healed by coalescence if and only if the two recombinants coalesce with each other before either coalesces with one of the remaining $l-1$ lineages. The probability of this can be shown to be $(2/3)l^{-1}$ (Saunders *et al.* 1984; Wiuf and Donnelly 1999). The probability that the recombination event happened while there were l lineages, given that it did happen, is shown by Griffiths and Marjoram (1997) to be

$$\frac{1}{(l-1)\sum_{k=1}^{n-1} k^{-1}},$$

so the total probability is

$$p(n) = \sum_{l=2}^{n} \frac{2}{3l(l-1)\sum_{k=1}^{n-1} k^{-1}} = \frac{2(n-1)}{3n\sum_{k=1}^{n-1} k^{-1}},$$

as required.