Theor Appl Genet (1998) 97:857–864

E. A. Thompson · T. R. Meagher

Genetic linkage in the estimation of pairwise relationship

Received: 26 November 1997 / Accepted: 3 March 1998

Abstract New types of markers, such as RAPDs, microsatellite markers, AFLPs, and SNPs provide the opportunity to obtain information on individuals at multiple genetic loci across the genome. This increase in the number of marker loci has provided enhanced opportunities for statistical analysis of the genetic consequences of genealogical relationship among individuals. In place of the classical models, we can now investigate empirical multilocus segregation patterns. Linkage among loci decreases the precision of relationship estimation but permits additional dimensions of genome sharing to be explored. In this paper we consider the effect of linkage on the pattern of genome sharing among relatives who share (on average) 25% of their dipolid genomes using the empirical meioses giving rise to 58 gametophytes from a single maternal plant of the species *Pinus taeda* (loblolly pine). The genome sharing among relatives is quantified in terms of the linkage map of the markers.

Key words Halfsib gametophytes · Multilocus gene identity · RAPD markers · Segregation indicators · Variance of genome sharing

Introduction

With the increasing availability of a genomic array of random amplified polymorphic DNA (RAPD) markers (Williams et al. 1990), microsatellite markers (Murray et al. 1994), amplified fragment length polymorphism (AFLP) markers (Vos et al. 1995), and single nucleotide polymorphisms (SNPs) (Chee et al. 1996), analysis of data at a dense array of linked markers is possible in natural populations. In the estimation of relationships between individuals from genetic data, the availability of data at large numbers of loci provides new opportunities for increased resolution in the assessment of genealogical relationships. Such analytical tools have led to advances in the assessment of breeding systems (e.g. Adams and Birkes 1991; Devlin and Ellstrand 1990; Meagher 1986; Meagher and Thompson 1987), reproductive success (e.g. Primack and Kang 1989; Roeder et al. 1989; Smouse and Meagher 1994; Snow and Lewis 1993), and social structure (e.g. Burke and Bruford 1987; Gibbs et al. 1994; Packer et al. 1991) of natural populations as well as in the development of conservation strategies for rare and endangered species (e.g. Brock and White 1992; Chase et al. 1996; Geyer et al. 1993).

The degree of genetic relationship between a pair of individuals is defined by the marginal probability of gene identity by descent at a given locus, and hence is the expected proportion of genome shared, which can be estimated empirically. For precise estimates of the realized proportion, data at large numbers of loci are essential, but the number of unlinked loci is necessarily limited. On the one hand, linkage decreases information, in the sense that the positive correlations between gene identity by descent at linked loci leads to the proportion of genome shared having a higher variance than for a set of independently segregating but equally informative loci (Thompson 1986). Thus, a set of linked loci provide a less precise estimate of the degree of relationship than with a set of unlinked loci with the same allele frequencies.

On the other hand, data at linked loci provide information not available from independently segregating loci. Thompson (1986) discusses the example of

Communicated by P. M. A. Tigerstedt

E. A. Thompson (✉)
Department of Statistics, University of Washington,
Box 354322, Seattle, WA 98195-4322, USA
Fax: 206-685-7419
E-mail: thompson@stat.washington.edu

T. R. Meagher
Department of Ecology, Evolution, and Natural Resources,
Rutgers University, Piscataway, NJ 08855, USA

halfsibs, grandparent-grandchild, and aunt-niece pairwise relationships. These three pairwise relationships are indistinguishable on the basis of data at independently segregating loci, each providing at each locus a probability 0.5 that the two individuals share one gene identical by descent and 0.5 that they share no genes identical by descent. However, the three relationships have distinct consequences for data at linked loci, since each provides a different probability that the two relatives share one gene identical by descent at both of two linked loci. Thompson (1988) extends this analysis to a consideration of kinship measures at two and three linked loci, showing that there are relationships that have identical two-locus kinship for all recombination frequencies between the two loci but also have distinct three-locus kinship if the three loci are linked.

Even were information available on an entire genome, the information for relationship estimation would be bounded due to the genome's finite length. Donnelly (1983) developed a framework for the analysis of genome sharing in relatives of various types and genome extinction in sets of offspring. The framework uses the simple no-interference model of Haldane (1919), in which crossovers occur as a Poisson process rate 1/Morgan along a chromosome, independently in every segregation. This same framework was extended by Bickeböller and Thompson (1996) to analyses of the probability distributions of surviving genomes in sets of descendants. The same model has recently also been analyzed by Feingold (1993) and by Guo (1994) in the context of linkage detection for a complex trait using a genomic array of markers.

Haldane's model is only an approximation, and the second aspect of the availability of a dense array of polymorphic markers is that we are no longer limited to computations under such a model. Instead, we may use empirically observed meioses in our analyses. In this paper we use only a small set of 58 plant meioses classified at 232 RAPD markers, which were provided by Professor R. Sederoff and colleagues at the North Carolina State University, in order to illustrate our approach of using empirical meioses to investigate broader questions of genome sharing among relatives. However, CEPH data (www address: http.//www.ceph. fr/HomePage.html) provide similar data for humans, and the amount of such data is increasing rapidly. CEPH data have recently been used by Lamb et al. (1997) to estimate patterns of meiotic exchange.

The specific objectives of this paper are to: (1) develop a general framework for evaluating the patterns of genome shared among relatives using data at multiple linked loci; (2) measure the effect of linkage between discrete marker loci on the precision of estimation of degree of relationship from data at those loci; (3) apply this framework using the empirical meioses observed in a progeny array from *Pinus taeda*, and (4) test for differences between the predictions based on empirical meioses and those based on the theoretical model and estimated genetic map.

Methods

Estimation of relationship

From the genealogy to phenotypic data, there are three steps. First, Mendelian segregation probability laws give rise to a pattern of segregation indicators, $\mathrm{Y} = \{Y_{ij}\}$:

$$\begin{aligned} &Y_{ij} = 1 \text{ if allele at locus } i \text{ in meiosis } j \text{ is grandmaternal} \\ &Y_{ij} = -1 \text{ if allele at locus } i \text{ in meiosis } j \text{ is grandpaternal} \end{aligned} \tag{1}$$

(These binary segregation indicators may equally be taken as 0 and 1, but the ± 1 notation is more convenient for the current paper since then Y_{ij} has mean 0 and variance 1.) Second, the genealogical structure or relationship R converts Y into the patterns of gene identity by descdent, B(Y), specifying the underlying genes shared by observed individuals. Third, the probability laws of population genetics provide the probability of observed genotypic or phenotypic data, given these underlying patterns of gene identity by descent, and hence a likelihood for a relationship R given the observed data.

$$L(\textit{relationship } R) = P(\textit{data} \mid R) = \sum_{\mathrm{B}} P(\textit{data} \mid \mathrm{B}) P(\mathrm{B} \mid R) \tag{2}$$

where $P(\mathrm{B} \mid R) = \sum_{Y \in \mathscr{Y}(\mathrm{B}, R)} P(\mathrm{Y})$

and $\mathscr{Y}(\mathrm{B}, R)$ is the set of Y-values that give rise to gene identity pattern B under relationship R.

In most analyses of inference of relationship, attention has been focused on the probability of data, given underlying patterns of gene identity by descent, and the effect of population parameters such as allele frequencies on these probabilities. However, where data at multiple linked marker loci are used, the effect of the assumed probability model for the multilocus segregation patterns in each meiosis should also be considered. All data, even sequence data, are of identity-by-state; it is impossible to observe identity-by-descent. However, in some situations it is possible to observe segregation indicators Y. In this paper, we use empirical multilocus segregation patterns, instead of a theoretical model, to generate gene identity by descent patterns in various relationships. The results are compared to a non-interference model on the estimated genetic map.

In the estimation of relationship from genetic data at multiple loci, the proportion of genome shared contains the largest part of the information. Indeed, if the loci are unlinked, relationships of the same degree, such as halfsibs, grandparent-grandchild, or aunt-niece, are formally indistinguishable (Thompson 1986). If data are available at linked loci, there is additional information in the lengths and patterns of segments of genome shared (Browning 1998). However, gene identities by descent at linked loci are positively correlated, so that information about the degree of relationship is less given data on linked loci than if given data at an equivalent set of unlinked loci. Thus, in addition to investigating the effects of using empirical meioses on patterns of multilocus genome sharing among relatives, we will also develop a natural measure of the effect of linkage between discrete marker loci on the precision of estimation of degree of relationship from data at those loci.

To make our discussion of these issues specific, in the following sections we focus on the three simplest relationships of equal degree: grandparent-grandchild (G), halfsib (H), and aunt-niece (N). For ease of terminology, we consider relationships in the maternal line.

Empirical observation of segregation

Although gene identity by descent and the underlying segregation indicators are never directly observable, in some practical instances they almost are so. One such case is that of data on RAPD markers in gametophytes from a single maternal plant, where the markers are chosen to be ones segregating in the offspring, so the maternal plant is known to be heterozygous for each RAPD band. Such data form the basis for the pseudo-testcross approach (Grattapaglia and Sederoff 1994) that has been widely used in genetic map construction for long-lived species such as trees.

For such genetic markers, gene identity among the gametophytes is known. The observed data are

$X_{ij} = 1$ if band for locus i is present in gametophyte j

$X_{ij} = -1$ otherwise (3)

for $j = 1, \ldots, n$ the number of gametophytes, and $i = 1, \ldots, m$ the number of loci. If $X_{ij} = X_{ik}$ gametophytes j and k share genome identical by descent from their mother at marker locus i. Further, the segregation indicators (1) are:

$Y_{ij} = 1$ if allele at locus i in gametophyte j is grandmaternal

$Y_{ij} = -1$ if allele at locus i in gametophyte j is grandpaternal (4)

The variables Y_{ij} are not directly observed, since it is not known whether a given band is maternal or paternal in the maternal plant. However,

$$X_{ij} = Z_i Y_{ij}$$

where $Z_i = \pm 1$ as the band at locus i is maternal/paternal in the maternal plant. The probability that $Z_i = +1$ and hence $X_{ij} = Y_{ij}$ is 0.5, marginally for each locus i, but the gametophyte scores at linked loci provide information on maternal haplotypes. In the absence of segregation distortion, and given a maternal plant heterozygous for all bands scored, the expectation of each X_{ij} and of each Y_{ij} is 0, and each has variance 1.

Let r_{ik} be the recombination frequency and d_{ik} be the map distance between locus i and locus k. Then

$$r_{ik} = (1 - \exp(-2d_{ik}))/2$$

$$\text{and } P(Y_{ij} = Y_{ki}) = 1 - r_{ik} = (1 + \exp(-2d_{ik}))/2 \quad (5)$$

For a set of n_{ik} independent meioses scored at both locus i and locus k, the sample covariance in segregation between locus i and k is $T^*_{ik} = n_{ik}^{-1} \sum_{j=1}^{n_{ik}} Y_{ij} Y_{kj}$. Since segregations to different offspring are independent, the statistic T_{ik} is the average of independent terms. Hence

$$E(T^*_{ik}) = P(Y_{ij} = Y_{kj}) - P(Y_{ij} \neq Y_{kj}) = \exp(-2d_{ik}) \quad (6)$$

$$\text{var}(T^*_{ik}) = n_{ik}^{-1}(1 - \exp(-4d_{ik})) \quad (7)$$

For unlinked loci $\sqrt{n_{ik}}\, T^*_{ik}$ has mean 0 and variance 1. For $n_{ik} \geqslant 20$, the distribution is very close to a standard normal $N(0, 1)$.

Although the Y_{ij} and hence T^*_{ik} are unobservable, the observable $T_{ik} = n_{ik}^{-1} \sum_{j=1}^{n_{ik}} X_{ij} X_{kj}$ differs from T^*_{ik} only in that it has a randomly assigned sign $Z_i Z_k = \pm 1$, depending on whether the bands i and k are concordant (both paternal/maternal) or discordant in the maternal genotype. To avoid this question, in a preliminary view of the dependence among loci, we consider the absolute value of the statistic $|T_{ik}| = |T^*_{ik}|$.

The proportion of genome shared by relatives

Suppose gametophyte j is scored at m_j loci. Since our binary variables are ± 1, we score the difference in proportion of genome shared and not shared. Between grandmother and gametophyte j this observed difference in proportion is $W_j = m_j^{-1} \sum_{i=1}^{m_j} Y_{ij}$. Although the expectation of each W_j is 0, in summing over loci, dependence between loci due to linkage enters into the formulae for the variance of shared genome. Equations 5 and 6 give immediately

$$\text{var}(W_j) = \text{E}(W_j^2) = m_j^{-2} \left(\sum_{i=1}^{m_j} \text{E}(Y_{ij}^2) + 2 \sum_k \sum_{i<k} \text{E}(Y_{ij} Y_{kj}) \right)$$
$$= m_j^{-1} + 2 m_j^{-2} \sum_k \sum_{i<k} \exp(-2d_{ik}) \quad (8)$$

The second term is the excess variance due to linkage.

Suppose that gametophytes j and l are both scored at m_{jl} loci. Between the halfsib gametophytes j and l, the difference in proportion of genome shared and not shared is

$$V_{jl} = m_{jl}^{-1} \sum_{i=1}^{m_{jl}} X_{ij} X_{il} = m_{jl}^{-1} \sum_{i=1}^{m_{jl}} Y_{ij} Y_{il}$$

since $Z_i^2 \equiv 1$. Then also

$$P(X_{kj} = X_{kl} \,|\, X_{ij} = X_{il}) = r_{ik}^2 + (1 - r_{ik})^2 = (1 + \exp(-4d_{ik}))/2$$

Again, each V_{jl} has expectation zero, and thus

$$\text{var}(V_{jl}) = \text{E}(V_{jl}^2)$$
$$= m_{jl}^{-2} \left(\sum_{i=1}^{m_{jl}} \text{E}((Y_{ij} Y_{il})^2) + 2 \sum_k \sum_{i<k} \text{E}(Y_{ij} Y_{il} Y_{kj} Y_{kl}) \right)$$
$$= m_{jl}^{-1} + 2 m_{jl}^{-2} \sum_k \sum_{i<k} \exp(-4d_{ik}) \quad (9)$$

the second term again being the excess in variance of genome sharing due to linkage. For an aunt-niece pair, the term $\exp(-4d_{ik})$ is replaced by $(\exp(-4d_{ik}) + \exp(-6d_{ik}))/2$ (Table 1). For a set of s loci, equispaced at distance d, expression 8 becomes

$$\text{var}(W_j) = s^{-1} + \frac{2(\exp(-2(s-1)d) - s + (s-1)\exp(2d)}{s^2(\exp(2d) - 1)^2} \quad (10)$$

The analogous expression for halfsibs ($\text{var}(V_{jl})$) simply replaces d by $2d$.

Note also that for three distinct gametophytes j, l, and r scored at the same m loci

$$m^2 \text{E}(V_{jl} V_{rl}) = \sum_{i=1}^{m} \text{E}(Y_{ij} Y_{ir} Y_{il}^2) + 2 \sum_k \sum_{i<k} \text{E}(Y_{ij} Y_{il} Y_{kr} Y_{kl}) = 0 \quad (11)$$

Thus, despite the dependence between loci, independence of the segregations to different offspring ensures the zero covariance of V_{jl} and V_{rl} (for distinct segregations j, l, and r). Hence, the empirical variance of genome sharing across all pairs of gametophytes provides an unbiased estimate of the theoretical value.

Patterns of genome sharing

In addition to having a different variance in the proportion of genome shared, the three relationships halfsibs (H), grandmother-grandchld (G), and aunt-niece(N) show different patterns in their segments of shared genome. Each segregation from the mother gives rise to portions of grandpaternal and grandmaternal genome, switching from one to the other as a Poisson process rate 1/Morgan along the chromosome. Grandmother and granddaughter (G) share genome wherever the segregating genome from the mother is grandmaternal. Note that the granddaughter's maternal genome may derive from either of the two chromosomes of the grandmother, whereas for maternal halfsibs (H) only the maternal genome can be shared. For relationship H, we have the superposition of the two

Table 1 Probabilities of gene identity at two linked loci for the three unilateral pairwise relationships at which individuals share 25% of their genome (From Thompson 1986)

Relationship	Grandmother G	Half-sisters H	Aunt-niece N
k_1; single locus	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
$k_{1\cdot1}(r)$; conditional	$1-r$	$R=(r^2+(1-r)^2)$	$(1-r)R+r/2$
as a function of d	$(1+\exp(-2d))/2$	$(1+\exp(-4d))/2$	$(2+\exp(-4d)+\exp(-6d))/4$

Note: k_1 is the probability that the two relatives share one gene identical by descent (*ibd*) at any autosomal locus. $k_{1\cdot1}(r)$ is the probability of sharing one gene *ibd* at a second locus at recombination frequency r (or map distance d; $r=(1-\exp(-2d))/2$), conditional on having one gene *ibd* at the first locus

independent segregations from the mother, giving rise to portions of identical (both grandpaternal/grandmaternal) and non-identical segments in the halfsibs' maternal chromosomes, switching between the two as a Poisson process rate 2/Morgan. Thus, genome sharing in *H* has the same overall expectation as in *G*, but the segments are (on average) half the size and there are (on average) twice as many.

For the aunt-niece pair (*N*) the process is more complicated. Both chromosomes of the aunt are involved but only the maternal chromosome of the niece. The aunt and her (full) sister share both their paternal and maternal genomes independently in the manner of halfsib shared genomes (that is, shared/not switching as a Poisson process rate 2). The segregation from the mother to her daughter superposes a Poisson process rate 1 which gives rise to the process illustrated in Fig. 1. The eight-state, specification of the three Poisson processes is Markov, but the process of genome sharing between the niece and her aunt is not, being the amalgum of four states. For gene identity at a pair of linked loci, the probabilities were given by Thompson (1986) and are summarized in Table 1.

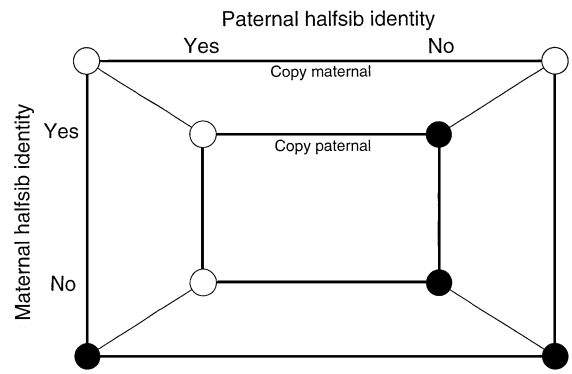

Fig. 1 Process of genome sharing in an aunt-niece pair. States are classified by whether or not the full sisters share the maternal /paternal genome and by the parental origin of the gene transmitted by one sister to her daughter. The four *white circles* correspond to states where the aunt and niece share a genome and the four *black ones* to where they do not

Application

The data

We illustrate the results of the previous section by application to 58 gametophytes from a single maternal plant of the species *Pinus taeda* (loblolly pine). The data were provided by Prof. R. Sederoff and colleagues at the North Carolina State University. Data are available on 232 RAPD markers for which the maternal plant is known to be heterozygous. There are few missing data, although not all gametophytes could be scored for all markers. Each of the 58 gametophytes is scored for at least 189 of the 232 loci (mean 219.6). One locus is scored in only 27 gametophytes, but the remainder are scored in at least 46 offspring (mean over all loci, 54.9). In all, the data are 92.7% complete (12,470 scorings out of a possible $13{,}456 = 232 \times 58$). Overall, there are 66 more scores of absent than of present, but no evidence of segregation distortion.

For the 232 loci, the sum over offspring ranges from -23 to $+19$, with a mean of -0.284, (SE = 0.950), not significantly different from 0. The expected range is -20.0 to $+20.0$. The empirical standard deviation of the 2-32 proportions

$$(presence - absence)/(number\ of\ offspring\ scored)$$

is 0.128, consistent with the theoretical value of 0.134, for a mean over an average of 54.9 independent segregations.

In the 58 individuals, the total over loci of (*presence* − *absence*) of bands ranges from -38 to $+33$ with a mean of -1.138 (SE = 1.079), not significantly different from 0. The expected range is -32.3 to $+32.3$. The empirical standard deviation of the 58 proportions

$$(presence - absence)/(number\ of\ loci\ scored)$$

is 0.073. For a mean over an average of 219.6 independent loci, the theoretical value for this standard deviation is 0.067. Thus, the realized effect of covariances induced by the common origins of bands at linked loci is only about 10%. (The realized covariances depend on the maternal haplotype).

Dependence among loci

Two markers are not only completely discordant in all the offspring scored but are scored in the same offspring. Six additional pairs, two sets of three, and one set of four loci are also not resolved by these data, being

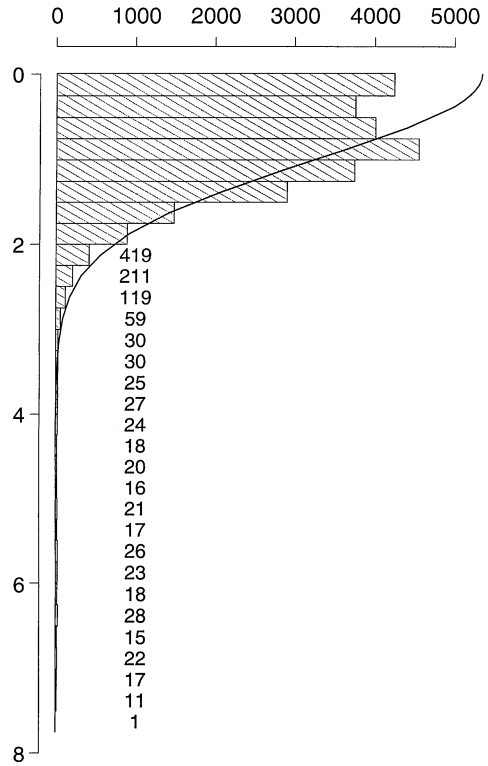

Fig. 2 Histogram of normalized pairwise relationship statistics between loci for all 26,796 pairs among the 232 RAPD loci, together with the standard Normal curve expected in the absence of linkage. Linkage creates extreme values between closely linked (or identical) loci, but the larger effect is in the deficiency of small values and excess of values in the range 1.5 to 2.0 which are the result of large numbers of pairs in loose linkage

fully concordant or fully discordant where scores are not missing. There are the same number of fully concordant as fully discordant pairs. These unresolved loci may in fact be the same locus, or closely linked loci not resolved by these data; in any case, the maximum likelihood estimate of recombination distance between them is zero. There are numerous other comparisons where only 1 or 2 of the 58 offspring show evidence of recombination, and many linkage groups can be distinguished and the loci within them ordered. Other loci show no clear linkage pattern. Our results agree with the map of Grattapaglia et al. (1991) which placed 191 of these 232 markers into 14 linkage groups.

The overall effect of linkage as measured by dependence among loci is given by the statistics $|T_{ik}^{*}|$ of Eq. 6. Figure 2 shows a histogram of these values for all 26,796 pairs of loci i and k, together with the corresponding null distribution. We see that linkage provides a few extreme values (the expected maximum in a sample of this size from the null distribution is about 4.0, while we have 276 pairs giving values larger than this). However, the major effect is between the much larger number of pairs in loose linkage, creating a deficiency of small values and a large excess in the range 1.5 to 2.0. While overall there is a strong effect of linkage, for many pairs of linked loci the value of T_{ik} would not indicate significant cosegregation.

Genome shared among relatives

These gametophytes directly illustrate genome sharing on the chromosomes of halfsibs deriving from their common parent. Note that although these halfsib haplotypes are dependent in their band presence/absence, all being the offspring of the same mother, they result from independent segregations and thus disjoint pairs are independent in their genome sharing. This fact is used repeatedly below in assessing the significance of genome-sharing patterns. The empirical variance of genome sharing across the markers (see Eq. 9) is 1/83.9. Thus, in terms of halfsib relationships, the 232 loci are equivalent in information content to about 84 independently segregating loci.

Since the complete linkage maps and maternal haplotypes are not easily determined for the full set of 232 markers, for the remainder of our analysis we consider only the two best-defined linkage groups of framework markers (Grattapaglia et al. 1991; O'Malley et al. 1996; R. Sederoff, personal communication). One group has $m_1 = 12$ markers over 125 cM, and the other has $m_2 = 11$ markers over 117 cM; these groups correspond to linkage groups 5 and 6, respectively, in O'Malley et al. (1996). To reduce missing data, we use the information on linked non-framework markers and the map to impute missing values of the framework markers. When this is done, almost all gametes can be scored with near certainty for almost all of the 23 loci. Additionally the two maternal haplotypes within each linkage group are clearly determined, so that for each locus set $l = 1, 2$, each gametophyte j provides an independent set of segregation indicators $\{Y_{ij}; i = 1, \ldots, m_l\}$ (Eq. 2) determined up to a single sign. This sign ambiguity does not affect the empirical variance of genome sharing; within each linkage group one maternal haplotype was arbitrarily designated the grandmaternal one.

Using the 58 gametes, we computed the variance of grandmaternal genome sharing for each linkage group. The variance of halfsib sharing was computed by comparing gamete j with gamete $j + 1$, gamete "59" being a duplicate of gamete "1". Although each gamete is involved in two comparisons, these are uncorrelated (see above). Between aunt and niece, there are five relevant segregations (Fig. 1). Each of the 58 gametes serves once as the maternal gamete of the "niece". The full-sister "mother"and "aunt" are created by a random choice of 4 other gametes, these 232 (58×4) choices being constrained by each gamete being represented the same number of times, no gamete serving twice in the construction of a given "aunt-niece" pair and no pair of gametes appearing together more than once in

Table 2 Variances of genome sharing, among 58 instances of each of three relationship types, for each of two linkage groups

	Locus set 1; 12 loci			Locus set 2; 11 loci		
	Grandmother	Half-sister	Aunt	Grandmother	Half-sister	Aunt
Mean	− 0.06	0.03	− 0.08	0.05	0.04	0.09
Standard error	0.08	0.07	0.07	0.08	0.07	0.06
Variance	0.415	0.282	0.273	0.383	0.269	0.236
Standard error	0.078	0.053	0.051	0.077	0.050	0.044
Variance excess due to linkage						
Observed	0.332	0.198	0.189	0.292	0.178	0.145
Expected	0.414	0.234	0.192	0.435	0.255	0.212
Effective loci	2.4	3.6	3.7	2.6	3.7	4.2

any of the sets giving rise to any "aunt-niece" pair. Thus, although each gamete is used five times in all, we obtain a set of 58 "aunt-niece pairs" whose genome sharing is uncorrelated.

Table 2 shows the results of these computations for each pairwise relationship and each group of loci. Given are the mean and standard deviation of the difference in genome shared and not-shared: These means have theoretical expectation zero, and none deviates significantly from this expectation. Given also are the empirical variances and their estimated standard errors, and two interpretations of these variances. First the empirical excess in variance over that for independently segregating loci is given, together with the theoretically expected values for this excess (the last terms of Eqs. 8 and 9). Finally, the inverse of the variance is a measure of the equivalent number of independently segregating loci. These results are discussed below.

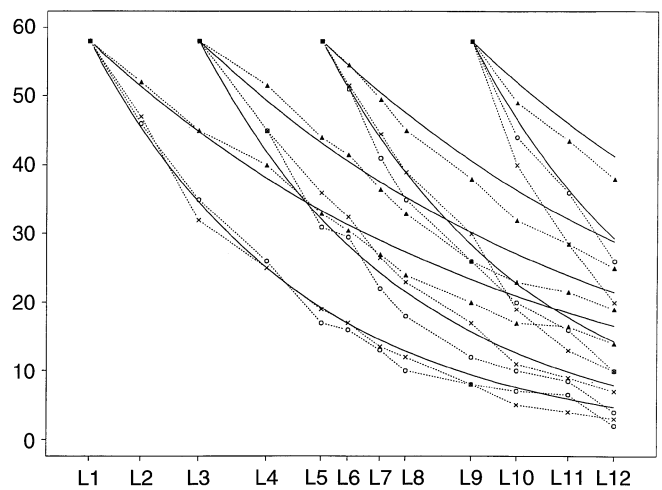

Fig. 3 Genome sharing in various relationships over the first linkage group of 12 loci stretching over 1.251 Morgans. The *points* show the decreasing numbers of pairs of individuals in a given relationship sharing genes over all loci typed from the "left-most" locus on a given curve to a given locus, conditional on sharing at this left-most locus. See text for details of the construction of these counts. The points for grandmothers are denoted by *triangles*, for half-sisters y *open circles*, and for aunt-niece by an $\times$. The *solid lines* give the theoretical expectations for the first two relationships (grandmothers and half-sisters), conditional on the estimated genetic map. The loci on the *horizontal axis* are positioned in accordance with this map

Patterns of genome sharing

We consider finally the pattern of genome sharing across a linkage group among relatives of various types. In analyzing genome sharing, we make use of the sign ambiguity in the segregation indicators $\{Y_{ij}; i = 1, \ldots, m_l\}$ to condition on sharing at any given locus, using either a gamete or its reverse image. For example, in assessing grandmaternal genome sharing conditional on sharing at locus L_i, we designate the maternal haplotype shared by the gametophyte at L_i as the "grandmaternal" one for the purposes of that comparison.

First, for each of the 58 gametophytes, for each locus L_i in turn, we compute the number that have the same maternal haplotype at locus L_{i+1}, then of these also at L_{i+2}, and so on, in order to obtain patterns of grandmaternal allele sharing over successive loci, conditional on have a grandmaternal allele at locus L_i. Consider, for example, the points designated by triangles in the set of lines starting at L_3 in Fig. 3. All 58 gametes are scored at L_3, and of these 51.5 show the same maternal haplotype at L_4 as at L_3 (the half-counts being due to missing data imputed proportionately). Then 44 of these have the same also at L_5, with successive counts of 41.5, 36.5, 33, and 26, at loci L_6 to L_9. Note loci L_5 to L_8 are very closely linked, with correspondingly small decreases in the count of gametes continuing to share the maternal haplotype over these loci, while the greater distance from L_8 to L_9 provides more opportunity for recombination. For clarity, Fig. 3 shows only the curves initiating at loci L_1, L_3, L_5, and L_9. Figure 4 shows those initiating at M_1, M_4, and M_8 only. The upper solid line in each set of curves is the theoretical

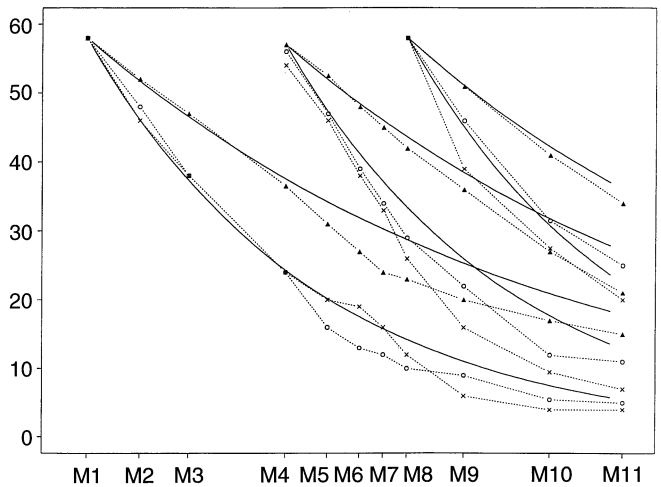

Fig. 4 Genome sharing in various relationships over the second linkage group of 11 loci stretching over 1.178 Morgans. The notation is as in Fig. 3

expectation of these observed counts, given the map distances of O'Malley et al. (1996).

As above, to assess half-sister (H) genome sharing, we compare each gamete j with gamete $j + 1$, gamete "59" being a duplicate of gamete "1". In comparing two different gametophytes, we condition on sharing genome at L_i by taking the reverse image of one gametophyte if necessary. As for the relationship G, we then plot continuing sharing over successive loci L_{i+1}, L_{i+2} and so on, as before, giving the open circles in Figs. 3 and 4. The lower solid curve again provides the theoretical expectation.

To demonstrate genome sharing between aunt and niece, we use each gamete five times in all to construct 58 "aunt-niece" pairs whose genome sharing is uncorrelated, as described above. From these pairs, and taking the mirror image of the niece's maternal gamete where necessary to condition on sharing at locus L_i, we again obtain the pattern of continuing genome sharing to locus L_{i+1}, L_{i+2}, and so on. These counts are denoted by $\times$ in Figs. 3 and 4.

Discussion

Among relatives that share the same overall proportion of genome identical by descent, a larger number of segregations in the descent paths defining the relationships provides more opportunities for recombination, and hence the shared genome is fragmented into smaller segments. Thus, there is no "equivalent number of independently segregating loci" corresponding to a length of genome. The number of independently segregating loci that will give an "equivalent" result depends on the function under consideration and on the relationship between individuals (Donnelly 1983). In the current paper we have seen that with respect to patterns and variance of genome sharing among halfsibs and grandparents, 11 or 12 loci in a linkage group of length about 1 Morgan are "equivalent" to a number of independently segregating loci ranging from 2.5 to 4 (Table 2).

In the estimation of degree of relationship (for example a kinship coefficient coefficient) between individuals from fully informative genetic data, the precision of estimation is inversely proportional to the variance of genome shared. For a given degree of kinship, the more complex the relationship the more precisely kinship can be estimated from data at a given set of linked loci, since the more opportunities for recombination results in a lesser degree of dependence in genome sharing at loci at given genetic distance. Conversely, in linkage analysis, it is well-recognized that more segregations provide more information for resolving very tightly linked loci; different relationships provide linkage information at different scales. For loose linkage ($0.05 \leqslant r \leqslant 0.2$ say), data on 58 fully informative segregations, such as the gametophytes of this paper or typical of some human pedigree studies, provide sufficient power both to detect linkage and to resolve loci. For tightly linked loci to be resolved, the net results of far more segregations are needed, leading to the use of recombinant inbred lines in experimental populations (Taylor 1978) or disequilibrium mapping in natural populations (Kaplan et al. 1995).

The curves of Figs. 3 and 4 and the results of Table 2 show good agreement with theory developed under Haldane's mapping function. However, in Table 2 it can be seen that each of the six excess variance figures is below its theoretical expectation, although none of the deficits is significant. In Figs. 3 and 4 (particularly the latter) it can also be seen that the data points tend to lie below the theoretical curves, likewise indicating greater-than-predicted breakup of the chromosome in these 58 segregations. This result should not be overinterpreted; the results do not deviate significantly from theory and, although our 58 comparisons were constructed to be uncorrelated within each relationship, the same 58 gametes are used for each of the three relationships considered.

If the result of a deficit in the excess variance of genome sharing due to linkage were to be confirmed by additional data, caution in interpretation is still needed. Significant effects of interference could be inferred, but these effects are confounded with map estimates. A better fit in Figs. 3 and 4 could be achieved by adjusting the map distances between the loci. The map of the framework markers used in this paper was constructed from these same data (Grattapaglia et al. 1991; O'Malley et al. 1996), but map estimation also assumes absence of interference, and the primary information for a linkage map derives from adjacent (but resolvable) loci. At close loci the curves of Figs. 3 and 4 provide an excellent fit. Stretching the map to obtain a better fit at

intermediate linkage distances (50–100 cM) would lead to an observation of too few recombination events at small distances (10–20 cM) relative to the theoretical no-interference model (see also Carter and Falconer 1951). In natural populations, it is hard to obtain sufficient data to detect interference (Bishop and Thompson 1988), and even more data will be needed to examine the effects of interference on patterns of genome sharing.

Acknowledgments The work was initiated while E. A. T was visiting Rutgers University; the hospitality of the Department of Biological Sciences is gratefully acknowledged. Professor Ron Sederoff (HNCSU) generously provided the RAPD marker data and updated map information used to illustrate the theory. This research was partially supported by NSF grant BIR 9305835.

References

Adams WT, Birkes DS (1991) Estimating mating patterns in forest tree populations. In: Fineschi S, Malvolti ME, Cannata F, Hattemer HH (eds) Biochemical markers in the population genetics of forest trees. SBP Academic, The Hague, pp 157–172

Bickeböller, H. Thompson EA (1996) The probability distribution of the amount of an individual's genome surviving to the following generation. Genetics 143:1043–1049

Bishop DT, Thompson EA (1988) Linkage information and bias in the presence of interference. Genet Epidemiol 5:107–120

Brock K, White BN (1992) Application of DNA fingerprinting to the recovery program of the endangered Puerto Rican parrot. Proc Natl Acad Sci 89:11121–11125

Browning SG (1998) Relationship information contained in gamete identity by descent data. Journal of Computational Biology 5:323–334

Burke T, Bruford M (1987) DNA fingerprinting in birds. Nature 327:149–152

Carter TC, Falconer DS (1951) Stocks for detecting linkage in the mouse, and the theory of their design. J Genet 50:307–323

Chase M, Kesseli R, Bawa K (1996) Microsatellite markers for population and conservation genetics of tropical trees. Am J Bot 83:51–57

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SPA (1996) Accessing genetic information with high-density DNA arrays. Science 274:610–613

Devlin B, Ellstrand NC (1990) The development and application of a refined method for estimating gene flow from angiosperm paternity analysis. Evolution 44:248–259

Donnelly KP (1983) The probability that related individuals share some section of the genome identical by descent. Theoret Popul Biol 23:34–64

Feingold E (1993) Markov processes for modeling and analyzing a new genetic mapping method. J Appl Probability 30:766–779

Geyer CJ, Ryder OA, Chemnick LG, Thompson EA (1993) Analysis of relatedness in the California condors, from DNA fingerprints. Mol Biol Evol 10:571–589

Gibbs HL, Goldizen AW, Bullough C, Goldizen AR (1994) Parentage analysis of multimale social groups of Tasmanian native hens (*Trobonyx mortierii*) – genetic evidence for monogamy and polyandry. Behav Ecol Sociobiol 5:363–371

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptis urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137:1121–1137

Grattapaglia D, Wilcox P, Chaparro JX, O'Malley DM, McCord S, Whetten R, McIntyre L, Sederoff R (1991) A RAPD map of loblolly pine in 60 days. Intl. Soc. for Plant Molecular Biology Intl Congr Abstr 2224. Kluwer Academic Publishers, Dordrecht, The Netherlands

Guo SW (1994) Computation of identity by descent proportions shared by two siblings. Am J Hum Genet 54:1104–1109

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Lamb NE, Feingold E, Sherman S (1997) Estimating meiotic exchange patterns from recombination data: an application in humans. Genetics 146:1011–1017

Meagher TR (1986) Analysis of paternity within a natural population of *Chamaelirum luteum*. I. Identification of most-likely parents. Am Nat 128:199–215

Meagher TR, Thompson EA (1987) Analysis of parentage for naturally established seedlings of *Chamaelirium luteum*. Ecology 68:803–812

Murray JC, Buetow KH, Weber JL, Ludwigson S, Scherpier-Heddema T, Manion F, Quillen J, Sheffield VC, Sunden S, Duyk GM, Weissenbach J, Gyapay G, Dib C, Morrissette J, Lathrop GM, Vignal A, White R, Matsunami N, Gerken S, Melis R, Albertsen H, Plaetke R, Odelberg S, Ward D, Dausset J, Cohen D, Cann H (1994) A comprehensive human linkage map with centimorgan density. Science 265:2049–2064

O'Malley DM, Grattapaglia D, Chaparro JX, Wilcox PL, Amerson HV, Liu B-H, Whetten R, McKeand S, Kuhlman EG, McCord S, Crane B, Sederoff R (1996) Molecular markers, forest genetics, and tree breeding. In: Perry Gustafson, Flavell RB (eds) 21st Stadler Genetics Symposium. Plenum Press, New York, pp 87–102

Packer C, Gilbert DA, Pusey AE, O'Brien SJ (1991) A molecular-genetic analysis of kinship and cooperation in African lions. Nature 351:562–565

Primack RB, Kang H (1989) Measuring fitness and natural selection in wild plant populations. Annu Rev Ecol Syst 20:367–396

Roeder K, Devlin B, Lindsay GB (1989) Applications of maximum likelihood methods to population genetic data for the estimation of individual fertilities. Biometrics 45:363–379

Smouse PE, Meagher TR (1994) Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (*Liliaceae*). Genetics 136:313–322

Snow AA, Lewis PO (1993) Reproductive traits and male fertility in plants: empirical approaches. Ann Rev Ecol Syst 24:331–351

Taylor BA (1978) Recombinant inbred strains: use in gene mapping. In: Morse HC III (ed) Origins of inbred mice. Academic Press, New York, pp. 423–428

Thompson EA (1986) Pedigree analysis in human genetics. The Johns Hopkins University Press, Baltimore

Thompson EA (1988) Two-locus and three-locus gene identity by descent in pedigrees. IMAJ Math Appl Med Biol 5:261–280

Vos P, Hogers R, Bleeker M, Reijans M, van der Lee T, Hornes M, Fritjers A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new concept for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Williams JGK, Kubelik AR, Livak KL, Rafalski JA, Tingey SR (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535