The Probability Distribution of the Amount of an Individual's Genome Surviving to the Following Generation

Heike Bickeböller* and Elizabeth A. Thompson†

** INSERM U.155, 75016 Paris, France and † Department of Statistics, University of Washington, Seattle, Washington 98195*

Manuscript received October 11, 1995
Accepted for publication March 7, 1996

ABSTRACT

The probability that at least $p\%$ of an individual's genome is passed on collectively to his children is calculated. With data availability the consideration of the chromosome as a whole rather than discrete loci becomes of increasing practical importance. Assuming the genomic continuum model, which allows for recombination, the crossover process in a chromosome pedigree is viewed as a continuous-time Markov random walk on the vertices of a hypercube with time parameter map distance along the chromosome. The desired probability corresponds to the probability of sojourn times of the process in a small set of vertices, which are well approximated via the Poisson clumping heuristic. Results are given for the human genome. It is very likely that an individual with at least four children passes on at least 90% of his genome. There exists no "equivalent" number of independently segregating loci for this distribution.

THIS article addresses a problem in pedigree analysis using a model for the locations of recombination events throughout the genome. Computing probabilities on pedigrees is important for many questions related to past, current and future generations. Most researchers consider discrete loci: one locus, a very small number of linked loci, or a finite number of unlinked loci. There are two reasons for this. First, the computational complexity of many problems increases rapidly with each additional locus. Second, in the past, most practical biological questions related to only a few loci.

Now data are available at densely packed loci; DNA sequencing is possible. Therefore models that consider chromosomes as a whole become of increasing importance. One might be interested in the ancestral path of a complete chromosome section. Relationships between individuals could be inferred more precisely if complete chromosomes rather than a few discrete loci could be compared. For the survival of rare species, genome rather than gene survival is ultimately important. A continuous model for chromosomes appears adequate for most purposes; the number of nucleotide bases is large and spaces between them are small.

Results yielded by analyses based on genomic continuum *vs.* discrete loci can be expected to be quite different, as shown by FISHER (1949), FRANKLIN (1977) and DONNELLY (1983). FISHER (1949) points out the difference between rates of approach to homozygosity at single loci and rates of approach to genomic homozygosity. FRANKLIN (1977) shows that individuals with the same inbreeding coefficient, *i.e.,* the same expected proportion of genome homozygous by descent (HBD), can have a different variance in the proportion of genome HBD. He discusses implications for the interpretation of population mean fitness as a function of population mean inbreeding, finding it important to consider the distribution, not only the expectation, of proportion of genome HBD.

DONNELLY (1983) considers the probability that two individuals in a given relationship share any (*i.e.,* not "no") genome identical by descent (IBD). This is a first result for determining the relationship between individuals by considering chromosome segments. He finds there is no "equivalent number" of independently segregating genes giving the same results as the recombination model with a genome of given length.

These considerations are relevant to an understanding of the role of homozygosity, inbreeding and gene survival in the genetic structure and fitness of a small inbred population, not just for single genes, but for chromosomes and the entire genome. Considering a genomic continuum will also be important in practical questions. Recent articles consider the identification of IBD chromosome segments in affected relatives (GOLDGAR 1990; FEINGOLD 1993; GUO 1994) or of homozygosity in affected children from consanguineous marriages (THOMPSON 1994) to localize a disease gene.

Of considerable interest for the maintenance of genome in rare species is the probability distribution of how much of an individual's genome is present in at least one descendant at a given generation. DONNELLY (1983) computed the probability of conserving the whole genome in c children. The objective in this article

Corresponding author: Heike Bickeböller, Institut für Medizinische Statistik und Epidemiologie, Klinikum r.d. Isar der Technischen Universität München, Ismaninger Str. 22, D-81675 München, Germany. E-mail: heike.bickeboeller@imse.med.tu-muenchen.de

is to compute *the probability that an individual passes on at least p% of his genome collectively to his offspring.* The underlying genomic continuum model is described below. The desired probability distribution can be derived directly from the distribution of genome shared IBD by c half-sibs, which we calculated previously via an approximation method (BICKEBÖLLER and THOMPSON 1996). The results yielded by the continuous model will be compared to those yielded by independently segregating discrete loci.

THE GENOMIC CONTINUUM MODEL

Chromosomes as a genomic continuum: The simplest model considering chromosomes as a continuum and allowing for recombination is due to HALDANE (1919) and FISHER (1949). If chromosomal distances are measured in Morgans (HALDANE 1919), crossovers occur as a Poisson process with rate 1. This does not allow for interference. However, the Poisson distribution is very precise if one considers large chromosome segments of the order of half a chromosome length. For many purposes it provides a good approximation to reality even for smaller distances and has been used in many contexts (FISHER 1949; DONNELLY 1983; LATHROP *et al.* 1985; GOLDGAR 1990; THOMPSON 1994).

The two chromosomes of a parent, represented by line segments, yield the chromosome that is passed on to the offspring. A crossover can be represented by breaking up and resplicing the two line segments involved at the crossover point (FISHER 1949). It is assumed that complete information on crossover locations is available in all parts of the genome considered. Only autosomal homologous chromosomes are considered, and equal map lengths for males and females are assumed.

Chromosome pedigrees and random walks: This framework is due to DONNELLY (1983). In *chromosome pedigrees* a nonfounder chromosome is obtained by a crossover process from the paternal and maternal chromosome of the parent, labeled 0 and 1, respectively. The crossover process may be considered as a *continuous-time Markov random walk* on these states, with "time" being position along the chromosome.

Each child's chromosome is an independent realization of the random walk. The process of all crossovers for one parent and his c children can be translated into a random walk on the vertices of a c-dimensional hypercube. For example, for three children suppose that, at position 0, the first two children have the paternal and the third has the maternal chromosome. The state is 001. At position t_1 a crossover happens in the third child. The random walk moves to state 000 and so on, until the end of the chromosome is reached. In the random walk model only one coordinate of the state-vector can change at position t. The process is assumed to start in the equilibrium distribution, which is uniform over all vertices. The processes for each of the segregations are independent.

The event that not all of the genome of an individual's chromosome pair of length l has been passed on to the offspring corresponds to the random walk hitting the hitting set $\mathcal{H} = \{00\cdots0, 11\cdots1\}$ at any position t before position l. Regions of the genome that are conserved in the offspring generation are regions where the random walk is *not* in $\mathcal{H}$.

The discrete-time random walk: Crossovers occur as a Poisson process. First, consider the *discrete-time random walk* that is given by the jump chain with a jump at each crossover. This chain is fully determined by the vector of initial probabilities and the transition matrix given in BICKEBÖLLER and THOMPSON (1996), and the probabilities for the number of visits to $\mathcal{H}$ by the random walk can be computed for a chromosome. Then the exponential distribution of intervals between crossover points provides the actual length of the random walk spent in $\mathcal{H}$.

PROBABILITY OF CONSERVING *P*% GENOME IN *C* CHILDREN

Use of the Poisson clumping heuristic: DONNELLY (1983) calculated the exact probability that an individual passes on 100% of his genome collectively to his c offspring, $p^c_{100\%}$, by computing first hitting times in $\mathcal{H}$. This is considerably easier than computing occupation times of the process in $\mathcal{H}$. These occupation times can be approximated by the Poisson clumping heuristic (ALDOUS 1989), although exact theoretical results cannot be found. With this approximation we previously computed the distribution of genome shared IBD by c half-sibs (BICKEBÖLLER and THOMPSON 1996). Regions of the genome in $\mathcal{H}$ correspond not only to regions of the genome *not* conserved but also to regions of the genome shared IBD by the c offspring.

Visits to $\mathcal{H}$ happen in widely spaced clumps where the clump centers follow a Poisson process. The heuristic is applied to the discrete-time random walk, approximating visits to $\mathcal{H}$ by the *union of independent and identically distributed (iid) random clusters with random centers chosen according to a Bernoulli process on* $\mathbf{Z}$ *(the set of all integers) with success rate* λ. Occupation times in $\mathcal{H}$ during length l are approximated by the union of those clumps whose centers lie in a window W that corresponds to the length l. The parameters for the approximation are given in the APPENDIX. The details are given by BICKEBÖLLER and THOMPSON (1996), who also demonstrate via simulations that the approximation works well. ALDOUS and BROWN (1993) give some recent theoretical results on the approximation of occupation times. Below we give the results for the distributions of interest here.

100% genome conserved in *c* children: Here results can be given in closed form. Let N_c denote the total number of autosomal chromosome pairs i, $i = 1, \ldots,$

TABLE 1

Probability of passing on 100% of an individual's genome for different number of children c

c	$p^{e}_{100\%}$	$p^{a}_{100\%}$	$p^{as}_{100\%}$
5	0.0000	0.0001	0.0000
6	0.0017	0.0026	0.0021
7	0.0280	0.0306	0.0271
8	0.1347	0.1356	0.1271
9	0.3268	0.3241	0.3134
10	0.5381	0.5390	0.5249
13	0.9038	0.9035	0.9006
20	0.9988	0.9988	0.9987

$p^{e}_{100\%}$, exact; $p^{a}_{100\%}$, approximation; $p^{as}_{100\%}$, simplified approximation.

N_c. Let L denote the total genome length. The probability of passing on 100% of an individual's genome is given by the product over all N_c chromosomes of the probability of no hits of $\mathcal{H}$ inside of the window for chromosome i. The formula for the Poisson clumping approximation, $p^{a}_{100\%}$, is given in the APPENDIX (Equation A5). There it is further approximated by

$$p^{as}_{100\%} = \exp\left(-\frac{cL}{2^{c-1}}\right). \qquad (1)$$

p% genome conserved in c children: Let Y denote the length (in M) that the random walk spends inside $\mathcal{H}$ during the total genome length L. Its distribution is given in the APPENDIX (Equation A6). It does not depend on the lengths of individual chromosomes. For length Y one allele of a particular locus is conserved. For length $(L - Y)$ both alleles are conserved. Let Z denote the percentage genome conserved in c children. Then

$$Z = \frac{1 \cdot Y + 2 \cdot (L - Y)}{2L} = 1 - \frac{Y}{2L}. \qquad (2)$$

The expectation and variance of Z are given by (APPENDIX, above Equation A7):

$$E(Z) = 1 - \frac{1}{2^c} + \frac{1}{2^c} \cdot O(c^{-1})$$

$$\mathrm{Var}(Z) = \frac{\mu}{2(Lcp_T)^2}, \qquad (3)$$

where $1/p_T$ denotes the mean cluster size and μ the Poisson parameter for the number of clumps during L (APPENDIX, Equations A2, A4).

RESULTS: THE DISTRIBUTION OF GENOME CONSERVED

In this paper, humans are used as an example. Humans have $N_c = 22$ autosomal chromosome pairs, and the total map length is taken to be $L = 33$. The probabilities of passing on 100% of an individual's genome are given in Table 1 for different number of children c.

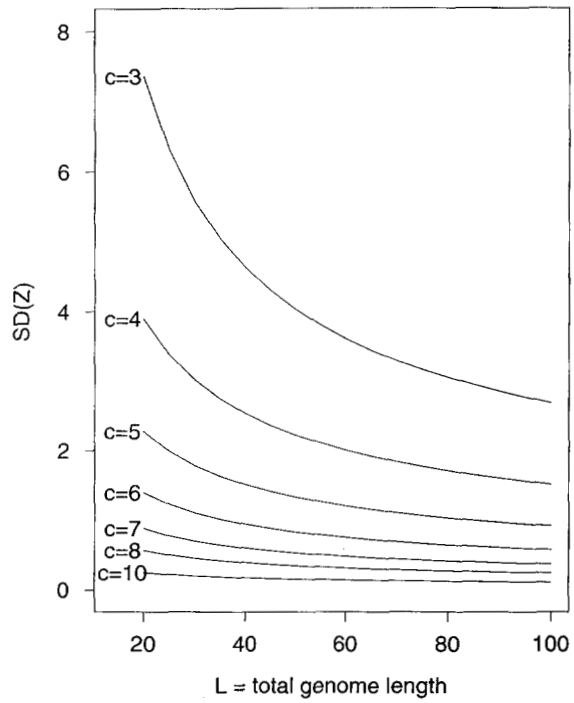

FIGURE 1.—Cumulative distribution function, $P(Z \leq z)$, of percentage genome conserved in c children for $3 \leq c \leq 10$ and $z \geq 90\%$.

The exact results, $p^{e}_{100\%}$, are discussed by DONNELLY (1983). With seven children, the chance of passing on all of an individual's genome is only 3%. To be 90% sure of passing on 100% genome at least 13 children are needed. These probabilities would decrease slightly if one also considers the sex chromosomes. The results for the Poisson clumping approximation, $p^{a}_{100\%}$, and the simplified approximation, $p^{as}_{100\%}$, agree well with the exact results.

The distribution of the proportion Z of an individual's genome conserved is computed with a FORTRAN program. Figure 1 shows the cumulative distribution function of the percentage genome conserved, $P(Z \leq z)$, for $3 \leq c \leq 10$ and $z \geq 90\%$. It is very likely for $c \geq 4$ that an individual passes on at least 90% of his genome. This is encouraging from a conservation biologist's point of view. Relaxing the requirement of passing on 100% to say 90–95% of an individual's genome gives a necessary number of children c that is at least feasable for most species.

COMPARISON WITH THE DISCRETE LOCI CASE

100% genome conserved in c children: The probability of passing on all of an individual's genome for N_l independently segregating loci is given by

$$\left(1 - \frac{1}{2^{c-1}}\right)^{N_l} \approx \exp\left(-\frac{N_l}{2^{c-1}}\right). \qquad (4)$$

Sensible choices for N_l seem to be 22 and 33, the number of autosome pairs and the map length of the total

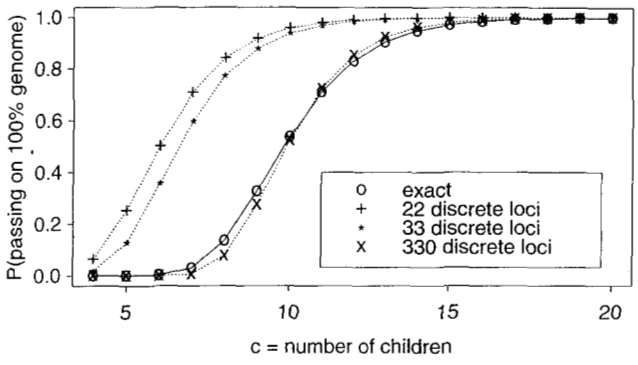

FIGURE 2.—Probability of passing on 100% of an individual's genome for different number of children.

genome, respectively. Figure 2 shows the probability of passing on 100% of an individual's genome as a function of c. The curves for 22 and 33 loci are quite different from the exact computation that allows for linkage. Linkage lowers considerably the probability of passing on 100% of an individual's genome.

A comparison of (1) and (4) shows that there is no number N_l of independent discrete loci that will give the same functional form (as a function of c) as yielded by considering recombination. However, for a particular c it is possible to find an equivalent number N_l, $N_l = cL$. Thus, for $c = 10$, $N_l = 330$. As shown in Figure 2, $N_l = 330$ gives a reasonable approximation to $p^c_{100\%}$ for all c. $N_l = 330$ is surprisingly high.

p% genome conserved in c children: Let Y_{N_l} denote the proportion of loci shared IBD by c half-sibs of a total of N_l independent loci. $N_l Y_{N_l}$ follows a binomial distribution with parameters N_l and $p = 1/(2^{c-1})$. Let Z_{N_l} denote the percentage genome conserved for N_l independent loci. In parallel to (2) and (3), the mean and variance of Z_{N_l} are given by

$$E(Z_{N_l}) = 1 - \frac{1}{2^c}$$

$$\mathrm{Var}(Z_{N_l}) = \frac{1}{2^{c+1}} \cdot \left(1 - \frac{1}{2^{c-1}}\right) \cdot \frac{1}{N_l}. \qquad (5)$$

The expectations of the percentage genome conserved considering N_l independently segregating loci and a continuous genome are the same. Comparing (3) and (5) shows the magnitude of the error in the approximation. Table 2 shows expectations and standard deviations. The columns "clumping theory", "clumping computation" and "discrete case" denote computations with the exact approximation formula (A7), with the FORTRAN program for the probability distribution and for the discrete loci, respectively. The expected values for the clumping heuristic and the discrete case agree very well, except for the cases $c = 3$ and $c = 4$, where the results are reasonably good although the clumping heuristic is not expected to work well.

Now consider the total distribution of Z. As above consider $N_l = 22$ and 33. Both distributions Z_{22} and Z_{33} are more spread than the distribution for Z (not shown). Differences between Z and Z_{22} are larger than between Z and Z_{33}. If one considers the probability of conserving at least p% of an individual's genome for decreasing p and different c, the differences between the genome continuum and 22 or 33 loci become smaller as one moves from the far right tail of the distribution of Z (not shown). Differences are noticeable for $c = 3$–5, values that are of practical interest. (The approximation for $c = 3$, 4 is not expected to be as good as for $c \geq 5$, see APPENDIX.) The effect of recombination *vs.* 22 or 33 loci is to reduce the probability of conserving 100% of the genome (Figure 2). This is *not* generally true for conserving at least p% of the genome.

Although $N_l = 330$ gives a good approximation of the probability of conserving 100% of an individual's genome for all c, there is no equivalent number N_l that will give the same probability for all c. $N_l = 330$ was found by computing the number of loci that gives the same results for $c = 10$ as the continuum using the (approximate) closed forms for the discrete and continuous case. This is not possible for general p%, since no closed form is available. One can, however, determine empirically the number of loci that best approximates the probability of conserving at least p%, as a function of c. This number N_l varies with p. For conserving at least 97, 95 and 90%, $N_l \approx 80$, $N_l \approx 80$–100 and $N_l \approx 70$–100, respectively, give good approximations. For smaller p%, there is a wide range of N_l that gives a reasonable approximation. However, this range does *not* include $N_l = 22$ or 33.

Another way of comparing the discrete and continuous case is to compute the number of loci, which yields the same variance as the continuous case for a specific number of children; *i.e.*, solve $\mathrm{Var}(Z) = \mathrm{Var}(Z_{N_l})$ for N_l as a function of c (Equations 3 and 5). Since each variance is four times the corresponding one for the proportion of genome shared IBD among c half-sibs (computed in BICKEBÖLLER and THOMPSON 1996), the same comparison holds. Thus for $c \geq 5$ the number of loci $N_l(c)$ yielding the same variance is approximated by

$$N_l(c) \approx \left(1 - \frac{1}{2^{c-1}}\right)(c-1)\frac{L}{2}.$$

For each c it is possible to find an equivalent $N_l(c)$, but the functional form $N_l(c)$ depends on which distributional properties are considered. For conserving 100% genome we have $N_l \approx cL$; for equal variances we have $N_l \approx cL/2$. Thus, there can be no general equivalence across c-values or across distributional properties.

RESULTS: ROBUSTNESS

The approximation for the distribution of percentage genome conserved by the Poisson clumping heuris-

TABLE 2

Expectations and standard deviations of Z (in percentages)

	$E(Z)$			$SD(Z)$	
c	Clumping theory	Clumping computation	Discrete case	Clumping theory	Clumping computation
3	81.94	87.26	87.50	5.23	5.64
4	92.01	92.85	93.75	2.84	2.55
5	96.25	96.53	96.88	1.69	1.57
6	98.19	98.31	98.44	1.05	1.00
7	99.12	99.18	99.22	0.67	0.64
8	99.57	99.59	99.61	0.43	0.43
9	99.79	99.80	99.80	0.28	0.28
10	99.89	99.91	99.90	0.19	0.18
13	99.99	99.99	99.99	0.06	0.06
20	100.00	100.00	100.00	0.004	0.003

tic depends on c, L and N_c. Relative chromosome lengths enter only in higher order corrections. The results are thus robust toward changes in the individual chromosome lengths while keeping the total length L constant. However, due to boundary effects in the approximation at chromosome ends, results are not robust to large changes in the number of chromosome pairs N_c. The heuristic should be valid as long as N_c is much smaller than cL (Equation A4).

The expectation of the percentage genome conserved in c children is independent of L. Any dependency is due to errors in the approximation. However, the variance of Z depends on L (Equation 3). Figure 3 shows the standard deviation of Z as a function of L for $3 \leq c \leq 10$ and $N_c = 22$. As cL increases $SD(Z)$ changes less with L. However, for small c and typical L for mammals ($L \approx 20 - 35$) the dependency on L is considerable.

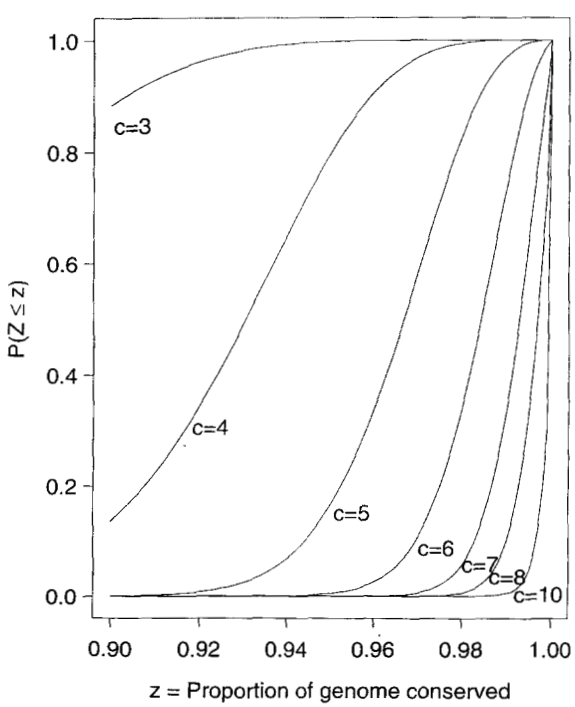

FIGURE 3.—Standard deviation of percentage genome conserved, $SD(Z)$, in c children for $3 \leq c \leq 10$ and total genome length $20 \leq L \leq 100$. Number of chromosomes $N_c = 22$ fixed.

The model assumes no interference, which changes the probability that in the vicinity of a crossover another crossover occurs. Thus, it changes the probability that a change in a coordinate of the state-vector of the random walk is immediately reversed. Hits to $\mathcal{H}$ still happen in widely spaced clusters with only a few hits of $\mathcal{H}$ per cluster. Thus the application of the Poisson clumping heuristic is valid. For positive interference a return to $\mathcal{H}$ in two steps is less likely. This leads to a decrease of the cluster sizes. The expected proportion of genome conserved should be increased and its variance reduced (Equations A7, 3). The "maximal" influence of positive interference can be suggested by assuming an expected cluster size of one. The effect is slight.

For negative interference a return to $\mathcal{H}$ in two steps is more likely. Thus negative interference will have opposite effects to positive interference. No simple bound can be given for negative interference, and as cluster sizes increase the Poisson clumping approximation may become less appropriate. However, negative interference should only be important in considering small chromosome sections.

DISCUSSION

We have considered the probability that an individual passes on at least $p\%$ of his genome collectively to his offspring. Although DONNELLY (1983) provided the result for $p = 100$, exact computation will seldom be possible. However, in conservation biology it is important to know how many offspring are required to have a high probability of conserving at least $p\%$ genome. The results are encouraging; it is very likely for $c \geq 4$ children that an individual passes on at least 90% of his genome.

The results here confirm the conclusions of FISHER (1949), FRANKLIN (1977) and DONNELLY (1983) for

the problems they considered. Results yielded by the genomic continuum and by independently segregating loci are considerably different; there is no "equivalent" number of independently segregating loci. The expectations of proportion of genome conserved are equal in the continuous and discrete case, but the distributions, including variances, are different. Differences are largest in the tail of the distribution, *i.e.*, for conserving almost the whole genome. For genome survival, interest is in this tail of the distribution, and it is thus essential to consider the genomic continuum model, which allows for recombination. While our results are not universal, they are reasonably robust to the assumptions and parameters used. The paper provides an approach that can be applied to other parameter choices.

We are grateful to DAVID ALDOUS, who encouraged the use of the Poisson clumping heuristic for this problem. This work was completed as part of the first author's Ph.D. thesis at the University of Washington, Seattle. The research has been funded by the National Science Foundation grants BSR-8921839 and BIR-9305835 and by a grant from the Graduate School Research Fund, University of Washington.

LITERATURE CITED

ALDOUS, D. J., 1989 *Probability Approximations via the Poisson Clumping Heuristic.* Springer, Berlin.

ALDOUS, D. J., and BROWN, M., 1993 Inequalities for rare events in time-reversible Markov chains II. Stochastic Process. Appl. **44:** 15–25.

BICKEBÖLLER, H., and E. A. THOMPSON, 1996 Distribution of genome shared IBD by half-sibs: approximation by the Poisson clumping heuristic. Theor. Pop. Biol. (in press).

DONNELLY, K. P., 1983 The probability that related individuals share some section of the genome identical by descent. Theor. Pop. Biol. **23:** 34–64.

FEINGOLD, E., 1993 Markov processes for modeling and analyzing a new genetic mapping method. J. Appl. Prob. **30:** 766–779.

FISHER, R. A., 1949 *The Theory of Inbreeding.* Oliver & Boyd, Edinburgh.

FRANKLIN, I. R., 1977 The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. Theor. Pop. Biol. **11:** 60–80.

GOLDGAR, D. E., 1990 Multipoint analysis of human quantitative genetic variation. Am. J. Hum. Genet. **47:** 957–967.

GUO, S. W., 1994 Computation of identity by descent proportions shared by two siblings. Am. J. Hum. Genet. **54:** 1104–1109.

HALDANE, J. B. S., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. **8:** 299–309.

LATHROP, G. M., J. M. LALOUEL, C. JULIER and J. OTT, 1985 Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am. J. Hum. Genet. **37:** 482–498.

THOMPSON, E. A., 1994 Monte Carlo estimation of multilocus autozygosity probabilities, pp. 498–506 in *Proceedings of the 1994 Interface Conference,* edited by J. SALL and A. LEHMAN. Interface Foundation of North America, Fairfax Station, VA.

Communicating editor: M. SLATKIN

APPENDIX

The parameters for the approximation: Details may be found in BICKEBÖLLER and THOMPSON (1996). Let $\mathcal{S}$ denote the union of iid distributed random clumps with random centers chosen according to a Bernoulli process on $\mathbf{Z}$ with success rate λ. Let W denote the window corresponding to a chromosome. $\mathcal{S} \cap W$ is approximated by the union of those clumps of $\mathcal{S}$ whose centers lie in W. $|\mathcal{S} \cap W|$ follows a compound Poisson distribution. A special case is the probability for the empty set:

$$P(\mathcal{S} \cap W \text{ empty}) \approx \exp(-\lambda |W|). \qquad \text{(A1)}$$

Two definitions of the cluster length are necessary. Let Γ^* denote the number of hits in a cluster and Γ^{**} the total number of steps in a cluster, including nonhits. A sensible cluster size distribution could be found for $c \geq 5$, but not for $c = 2, 3, 4$, due to the small number of orbits. A cluster is terminated if the random walk spends at least two steps outside of $\mathcal{H}$ for $5 \leq c \leq 9$ and at least four steps outside of $\mathcal{H}$ for $c \geq 10$.

For c children the expected cluster sizes $E(\Gamma^*)$ and $E(\Gamma^{**})$ are given as follows:

$$E(\Gamma^*) \equiv \frac{1}{p_T} = \begin{cases} \dfrac{c}{c-1} & \text{if } 5 \leq c \leq 9 \\ \dfrac{c^3}{c^3 - c^2 - 2c + 2} & \text{if } c \geq 10 \end{cases} \qquad \text{(A2)}$$

$$E(\Gamma^{**}) = \begin{cases} \dfrac{c+1}{c-1} & \text{if } 5 \leq c \leq 9 \\ \dfrac{c^3 + c^2 + 6c - 6}{c^3 - c^2 - 2c + 2} & \text{if } c \geq 10. \end{cases} \qquad \text{(A3)}$$

The equilibrium probability p of the process to be in $\mathcal{H}$, the rate λ for the position of the cluster center, the window size $|W|$ for chromosome length l and the Poisson parameter μ for the number of clumps during length L of the whole genome are given by

$$p = \frac{\text{No. of vertices in } \mathcal{H}}{2^c} = \frac{1}{2^{c-1}},$$

$$\lambda = \frac{p}{E(\Gamma^*)} = \frac{1}{E(\Gamma^*)\,2^{c-1}},$$

$$|W| = cl + E\Gamma^{**}, \quad \mu = \frac{cL + N_c E(\Gamma^{**})}{E(\Gamma^*)\,2^{c-1}}. \qquad \text{(A4)}$$

100% genome conserved in c children: For the probability of passing on 100% of an individual's genome, $p^a_{100\%}$, (A1) needs to be independently applied to each chromosome pair of length l_i, $i = 1, \cdots N_c$.

$$p^a_{100\%} = \prod_{i=1}^{N_c} \left(\exp\left(- \frac{cl_i + E(\Gamma^{**})}{E(\Gamma^*)\,2^{c-1}} \right) \right)$$

$$= \exp\left(- \frac{cL + N_c E(\Gamma^{**})}{E(\Gamma^*)\,2^{c-1}} \right). \qquad \text{(A5)}$$

An even simpler approximation, $p^{as}_{100\%}$, can be derived with (A2) and (A3). For $5 \leq c \leq 9$

$$p^{a}_{100\%} = \exp\left(-\frac{c(c-1)L + N_c(c-1)\left(1+\frac{2}{c-1}\right)}{c2^{c-1}}\right)$$

$$= \exp\left(-\frac{(c-1)L + N_c + O(c^{-1})}{2^{c-1}}\right)$$

$$= \exp\left(-\frac{cL + O(1)}{2^{c-1}}\right).$$

The same result holds for $c \geq 10$. Hence

$$p^{as}_{100\%} \equiv \exp\left(-\frac{cL}{2^{c-1}}\right)$$

is an approximation to $p^{a}_{100\%}$ and thus to $p^{e}_{100\%}$, the exact probability of passing on 100% of an individual's genome. A similar formula was found empirically by DONNELLY (1983).

The simplified approximation can also be derived by the following heuristic argument, since $\mathcal{H}$ consists of only one orbit and has only one neighboring orbit. Consider a continuous-time random walk with two states, states $\mathcal{H}$ and $\bar{\mathcal{H}}$. The transition probabilities are

$$P_{\mathcal{H}\to\bar{\mathcal{H}}} = c \quad \text{and} \quad P_{\bar{\mathcal{H}}\to\mathcal{H}} = \alpha.$$

With $\mathcal{H}$ given as the same hitting set considered so far, and using the properties of a two state random walk, we have

$$P(\mathcal{H}) = \frac{1}{2^{c-1}} \quad \text{and} \quad P(\mathcal{H}) = \frac{\alpha}{\alpha + c}.$$

Therefore,

$$\alpha \approx \frac{c}{2^{c-1}}.$$

Hence the hitting time until time L is up to first order approximation

$$\exp(-\alpha L) = \exp\left(-\frac{cL}{2^{c-1}}\right).$$

p% genome conserved in c children: The distribution of the length Y that the random walk spends inside $\mathcal{H}$ during L is given by

$$f(y) = e^{-\mu}\left[I_{\{0\}}(0) + \sum_{x=1}^{\infty}\sum_{m=1}^{x}\binom{x-1}{x-m} \times \frac{\mu^m}{m!}p_T^{m-1}q_T^{x-m}\frac{c^x y^{x-1}e^{-cy}}{(x-1)!}I_{\{y>0\}}(y)\right], \quad \text{(A6)}$$

where p_T and μ are given by (A2), (A4) and $q_T = 1 - p_T$.

The expectation and variance of Z follow from those of $Y^* = Y/L$ given in BICKEBÖLLER and THOMPSON (1996). We compute the expectation of Z further, using (A2), (A3), and (A4).

$$E(Z) = 1 - \frac{\mu}{2Lcp_T} = 1 - \frac{1}{2^c} - \frac{N_c E(\Gamma^{**})}{Lc2^c}$$

$$= 1 - \frac{1}{2^c} + \frac{O(c^{-1})}{2^c}. \quad \text{(A7)}$$