Distribution of Genome Shared Identical by Descent by Two Individuals in Grandparent-Type Relationship

Valeri T. Stefanov

Department of Mathematics and Statistics, University of Western Australia, Nedlands 6907, Western Australia, Australia

Manuscript received April 18, 1999
Accepted for publication July 3, 2000

ABSTRACT

A methodology is introduced for numerical evaluation, with any given accuracy, of the cumulative probabilities of the proportion of genome shared identical by descent (IBD) on chromosome segments by two individuals in a grandparent-type relationship. Programs are provided in the popular software package Maple for rapidly implementing such evaluations in the cases of grandchild-grandparent and great-grandchild–great-grandparent relationships. Our results can be used to identify chromosomal segments that may contain disease genes. Also, exact P values in significance testing for resemblance of either a grandparent with a grandchild or a great-grandparent with a great-grandchild can be calculated. The genomic continuum model, with Haldane's model for the crossover process, is assumed. This is the model that has been used recently in the genetics literature devoted to IBD calculations. Our methodology is based on viewing the model as a special exponential family and elaborating on recent research results for such families.

THE genes at a given locus of two related individuals are said to be identical by descent (IBD) if one is a physical copy of the other or both are physical copies of the same gene in a common ancestor. Calculations associated with the concept of IBD at a single locus or at a finite (small) number of (linked) loci have been published in the genetics literature since the early 1940s (*cf.* the references in Bickeboller and Thompson 1996a). Such calculations become difficult with the increase of the number of loci and/or relatives. Moreover, considering a finite, however large, number of independent loci cannot account for the possibility of recombination. The latter problem can be overcome by considering the chromosomes as a continuum and modeling the occurrence of crossovers by a point process (*cf.* Lange 1997, Chap. 12). The latter is usually a Poisson process. It does not account for the possibility of interference but provides a good approximation to reality if long chromosomal regions are considered. This model dates back to Haldane (1919) and Fisher (1949). There has recently been increased interest in IBD calculations for the genomic continuum model. This is mainly due to the availability of data on densely packed loci, which makes the concept of IBD sharing for chromosomal regions or the whole genome of practical importance. The first result concerning IBD calculations, in the framework of the continuum model, is due to Donnelly (1983). He calculates the probability that individuals in a given relationship share any part of the genome IBD.

Note that the proportion of genome-shared IBD by related individuals is a random variable unless we consider some trivial cases, such as twins or a parent with a child. Donnelly (1983) calculates the probability that this random variable is positive. Its distribution function is generally unknown. The first result, although not exact, concerning its distribution in the case of c half-sibs, is due to Bickeboller and Thompson (1996a; *cf.* Bickeboller and Thompson 1996b). They find approximations to it using the Poisson clumping heuristic. Note that no exact result for the distribution function of the aforementioned random variable is available even in the simplest case of two closely related individuals, such as half-sibs or a grandparent and a grandchild. Earlier results provide only the expected value and variance of this random variable and the conditional counterparts of these given information on flanking markers (see Goldgar 1990; Hill 1993; Guo 1994a,b, 1995; Thompson 1995).

Throughout the article only autosomal chromosome segments are considered. Equal map lengths are assumed for male and female (*cf.* Bickeboller and Thompson 1996a).

In this article we introduce a methodology for numerical evaluation, with any given accuracy, of the cumulative probabilities of the proportion of genome shared IBD on chromosome segments by two individuals in a grandparent-type relationship. We provide Maple V programs for implementing such evaluations in the cases of grandchild-grandparent and great-grandchild–great-grandparent relationships. Also, these evaluate the cumulative probabilities given any information (*e.g.*, such as inheritance) on one of the flanking markers.

Author e-mail: stefanov@maths.uwa.edu.au

These are the first exact distributional results concerning IBD calculations in the framework of the genomic continuum model. Our methodology is applicable to higher-order grandparent-type relationships at the expense of heavier computational effort. It is also applicable to other relationships as long as the associated underlying mathematical models (continuous time Markov chains) do not have too many states. Roughly speaking, the latter means that we consider a small number of closely related individuals—for example, several half-sibs. Our results can be used in identifying chromosomal segments that may contain disease genes. Also, exact P values can be derived in significance testing for resemblance of a grandparent with a grandchild and of a great-grandparent with a great-grandchild.

THE UNDERLYING MATHEMATICAL MODEL

Haldane (1919) and Fisher (1949) have suggested that chromosomes be considered as a continuum and that the occurrence of crossovers along the chromosomes be modeled by a Poisson process. If the distances are measured in morgans, then the rate of the Poisson process is one. Donnelly (1983) elaborated on this model and showed that all crossover processes on a pedigree can be viewed as a continuous time Markov chain, whose states are the vertices of a hypercube. Also, the genome-shared IBD by a group of related individuals equals the sojourn time at a set of vertices up to time d, where d is the length (in morgans) of the chromosome segment of interest. For example, the amount of genome inherited by a great-grandchild from a great-grandparent equals the sojourn time at the vertex (1, 1) in a continuous time Markov random walk on the four vertices (1, 1), (1, 0), (0, 1), and (0, 0) of the two-dimensional unit cube, where the holding times at all vertices are exponentially distributed with parameter 2 (*cf.* Donnelly 1983; Guo 1995, p. 1473). Likewise, the amount of genome inherited by a grandchild from a grandparent equals the sojourn time at state 1 in a continuous time Markov random walk on the two vertices 1 and 0 of the one-dimensional unit cube, where the holding times are exponentially distributed with parameter 1. More specifically, the model for the first relationship (great-grandchild–great-grandparent) is a four-state continuous time Markov chain whose parameters are described as follows. Denote the states (1, 1), (1, 0), (0, 1), and (0, 0) by 1, 2, 3, and 4, respectively. The holding times are exponentially distributed with parameter 2 and the one-step transition probability matrix of the embedded discrete time Markov chain is given by

$$\begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix}.$$

The initial probability vector is (¼, ¼, ¼, ¼) (the steady-state probabilities). The model for the second relationship (grandchild-grandparent) is a two-state continuous time Markov chain whose states 1 and 0 we denote by 1 and 2, respectively. The holding times are exponentially distributed with parameter 1 and the one-step transition probability matrix of the embedded discrete time Markov chain is given by

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The initial probability vector is (½, ½). Let d be the length (in morgans) of the chromosome segment of interest. The quantity of interest is the sojourn time in state 1 within time interval of length d. This quantity divided by d is the proportion of the genome shared IBD on that segment by the individuals in question.

METHODS

Our methodology is based on the following key points. First, the underlying model can be viewed as a member from a special exponential family. Second, recent research results on such families (*cf.* Stefanov 1991, 1995) are applicable to get explicit expressions for the characteristic functions of relevant stopping times. Third, these characteristic functions are numerically invertable using the system Maple V (Monagan *et al.* 1997) and some numerical tools. Therefore, their distribution functions are derivable. Fourth, the latter distribution functions yield the distribution function of a sojourn time in a state within any fixed time interval. Subsequently the cumulative probabilities of relevant proportions of genome-shared IBD can be calculated. More details follow.

Let $\{X(t)\}_{t\geq 0}$ be a continuous-time Markov chain with four states whose parameters are described below. The embedded discrete-time Markov chain has the one-step transition probability matrix

$$\begin{bmatrix} 0 & p_1 & q_1 & 0 \\ p_2 & 0 & 0 & q_2 \\ p_3 & 0 & 0 & q_3 \\ 0 & p_4 & q_4 & 0 \end{bmatrix},$$

where $0 < p_i, q_i < 1$, and $p_i + q_i = 1$. Also λ_1 is the parameter of the exponentially distributed holding time in state 1 while λ_2 is the same for the remaining three states. Note that, in particular, if $p_i = q_i = 0.5$ and $\lambda_1 = \lambda_2 = 2$ we get the underlying mathematical model for the relationship great-grandchild–great-grandparent. Let $N_{ij}(t)$ be the number of one-step transitions from state i to state j up to time t. Also let $T_i(t)$ be the sojourn time in state i up to time t. Denote

$$T(t) = T_1(t) \quad \text{and} \quad S(t) = T_2(t) + T_3(t) + T_4(t).$$

Then, given the initial state is fixed, the likelihood function of the chain *X*, observed up to time *t*, is given by

$$\begin{aligned}\exp\{&N_{12}(t)\ln(p_1\lambda_1) + N_{13}(t)\ln(q_1\lambda_1) + N_{21}(t)\ln(p_2\lambda_2)\\ &+ N_{24}(t)\ln(q_2\lambda_2) + N_{31}(t)\ln(p_3\lambda_2) + N_{34}(t)\ln(q_3\lambda_2)\\ &+ N_{42}(t)\ln(p_4\lambda_2) + N_{43}(t)\ln(q_4\lambda_2)\\ &- \lambda_1 T(t) - \lambda_2 S(t)\}\end{aligned} \tag{1}$$

(*cf.* STEFANOV 1991). Introduce the following stopping time:

$$\tau_s = \inf\{t : T(t) \geq s\}, \quad s > 0;$$

that is, τ_s is the waiting time till the sojourn in state 1 reaches the level *s*.

Recall that $T(t)$ (the sojourn time in state 1 up to time *t*) identifies the length of the genome shared by the two individuals on a chromosome segment of length *t*. It is easy to see that the distribution functions of $T(t)$ and τ_s are related as

$$P(T(t) \geq s) = P(\tau_s \leq t). \tag{2}$$

Also

$$\begin{aligned}P(\tau_s \leq t) &= \sum_{i=1}^{4} P(\tau_s \leq t | X(0) = i)P(X(0) = i)\\ &= \sum_{i=1}^{4} P(S(\tau_s) \leq t - s | X(0) = i)P(X(0) = i),\end{aligned} \tag{3}$$

because $\tau_s = T(\tau_s) + S(\tau_s) = s + S(\tau_s)$.

The following propositions provide explicit expressions for the characteristic functions of $S(\tau_s)$ corresponding to different initial states. Their proofs are found in the APPENDIX.

PROPOSITION 1. *Assume that* $p_i = q_i = 0.5$ *for each i,* $\lambda_1 = \lambda_2 = 2$, *and* $X(0) = 1$. *Then the characteristic function of* $S(\tau_s)$ *is given by*

$$M^{(1)}_{S(\tau_s)}(t) = \exp\left(-2s + \frac{2s(1 - It/2)}{2(1 - It/2)^2 - 1}\right),$$

where $I = \sqrt{-1}$.

PROPOSITION 2. *Assume that* $p_i = q_i = 0.5$ *for each i,* $\lambda_1 = \lambda_2 = 2$, *and either* $X(0) = 2$ or $X(0) = 3$. *Then the characteristic function of* $S(\tau_s)$ *is given by*

$$M^{(2)}_{S(\tau_s)}(t) = \frac{1 - It/2}{2(1 - It/2)^2 - 1}\exp\left(-2s + \frac{2s(1 - It/2)}{2(1 - It/2)^2 - 1}\right).$$

PROPOSITION 3. *Assume that* $p_i = q_i = 0.5$ *for each i,* $\lambda_1 = \lambda_2 = 2$, *and* $X(0) = 4$. *Then the characteristic function of* $S(\tau_s)$ *is given by*

$$M^{(4)}_{S(\tau_s)}(t) = \frac{1}{2(1 - It/2)^2 - 1}\exp\left(-2s + \frac{2s(1 - It/2)}{2(1 - It/2)^2 - 1}\right).$$

Consider now the underlying model for the relationship grandchild-grandparent, which has been introduced in the preceding section. The relevant likelihood function is

$$\exp\{N_{12}(t)\ln\lambda_1 + N_{21}(t)\ln\lambda_2 - \lambda_1 U_1(t) - \lambda_2 U_2(t)\},$$

where $N_{ij}(t)$ is the number of one-step transitions from state *i* to state *j* in the time interval $[0, t]$, and $U_i(t)$ is the sojourn time in state *i* in the same time interval. Similarly to the preceding case, introduce the stopping time

$$\nu_s = \inf\{t : U_1(t) \geq s\}, \quad s > 0,$$

and note that

$$P(U_1(t) \geq s) = P(\nu_s \leq t). \tag{4}$$

Recall that if $\lambda_1 = \lambda_2 = 1$ then $U_1(t)$ is the length of the genome inherited by a grandchild from his grandparent on a chromosome segment of length *t*. Also

$$\begin{aligned}P(\nu_s \leq t) &= \sum_{i=1}^{2} P(\nu_s \leq t | X(0) = i)P(X(0) = i)\\ &= \sum_{i=1}^{2} P(U_2(\nu_s) \leq t - s | X(0) = i)P(X(0) = i)\end{aligned} \tag{5}$$

(we use the same notation, *X*, for the corresponding two-state Markov chain). Note that $N_{12}(\nu_s) = N_{21}(\nu_s)$ if $X(0) = 1$ and $N_{12}(\nu_s) = N_{21}(\nu_s) - 1$ if $X(0) = 2$. The following hold.

PROPOSITION 4. *Assume that* $\lambda_1 = \lambda_2 = 1$ *and* $X(0) = 1$. *Then the characteristic function of* $U_2(\nu_s)$ *is given by*

$$M^{(1)}_{U_2(\nu_s)}(t) = \exp\left(\frac{Ist}{1 - It}\right).$$

PROPOSITION 5. *Assume that* $\lambda_1 = \lambda_2 = 1$ *and* $X(0) = 2$. *Then the characteristic function of* $U_2(\nu_s)$ *is given by*

$$M^{(2)}_{U_2(\nu_s)}(t) = \frac{1}{1 - It}\exp\left(\frac{Ist}{1 - It}\right).$$

The tools used for deriving the aforementioned propositions are applicable to the sojourn time in any given state from any given finite-state Markov chain (*cf.* the APPENDIX) and are therefore also applicable to higher-order grandparent-type relationships.

To compute the cumulative probabilities of the proportion of shared genome we need the conditional cumulative probabilities of $S(\tau_s)$ and $U_2(\nu_s)$ given the initial state [*cf.* the identities (2), (3), (4), and (5)]. The above propositions provide the characteristic functions of these conditional distributions. We invert them, using some numerical tools, and find the required cumulative probabilities. The mathematical details of these are provided in the APPENDIX.

RESULTS

We provide two Maple V programs in the APPENDIX. These evaluate the cumulative probabilities of the ge-

TABLE 1

Cumulative probabilities ($F_d(x)$) of the proportion (x) of genome-shared IBD on a chromosome segment of length d morgans by two individuals in the relationship grandchild-grandparent

x	$F_{0.5}(x)$	$F_{1.75}(x)$	$F_3(x)$
0.00	0.303265	0.086887	0.024894
0.05	0.322323	0.117050	0.046587
0.10	0.341573	0.150423	0.074591
0.12	0.349323	0.164630	0.087617
0.14	0.357098	0.179306	0.101695
0.16	0.364899	0.194435	0.116821
0.18	0.372723	0.210000	0.132985
0.20	0.380570	0.225982	0.150171
0.22	0.388437	0.242362	0.168356
0.24	0.396324	0.259118	0.187509
0.26	0.404229	0.276230	0.207594
0.28	0.412151	0.293673	0.228567
0.30	0.420088	0.311423	0.250378
0.32	0.428040	0.329456	0.272973
0.34	0.436003	0.347746	0.296289
0.36	0.443978	0.366266	0.320262
0.38	0.451963	0.384990	0.344819
0.40	0.459956	0.403888	0.369886
0.42	0.467957	0.422933	0.395384
0.44	0.475963	0.442096	0.421300
0.46	0.483973	0.461348	0.447339
0.48	0.491986	0.480659	0.473625
0.50	0.500000	0.500000	0.500000

TABLE 2

Cumulative probabilities ($F_d(x)$) of the proportion (x) of genome-shared IBD on a chromosome segment of length d morgans by two individuals in the relationship great-grandchild–great-grandparent

x	$F_{0.5}(x)$	$F_{1.75}(x)$	$F_3(x)$
0.00	0.547465	0.261424	0.125676
0.05	0.571825	0.330677	0.210284
0.10	0.595543	0.398552	0.299645
0.15	0.618608	0.464172	0.390096
0.20	0.641012	0.526809	0.478449
0.25	0.662746	0.585888	0.562118
0.30	0.683805	0.640974	0.639164
0.35	0.704185	0.691771	0.708297
0.40	0.723884	0.738107	0.768824
0.45	0.742900	0.779921	0.820573
0.50	0.761233	0.817247	0.863789
0.55	0.778887	0.850203	0.899039
0.60	0.795863	0.878975	0.927106
0.65	0.812166	0.903801	0.948896
0.70	0.827802	0.924958	0.965364
0.75	0.842777	0.942750	0.977451
0.80	0.857098	0.957497	0.986032
0.85	0.870775	0.969524	0.991895
0.90	0.883815	0.979153	0.995718
0.95	0.896230	0.986696	0.998062
0.99	0.905719	0.991429	0.999181

nome-shared IBD on chromosome segments by two individuals in either grandchild-grandparent or great-grandchild–great-grandparent relationship. It takes a few minutes real time to execute the longer program on either a PC or UNIX workstation. Tables 1 and 2 provide excerpts of such cumulative probabilities for chromosome segments of length 0.5, 1.75, and 3, respectively. Table 1 does not contain quantiles larger than the median because the distribution function for the grandchild-grandparent relationship is symmetric.

The user of these programs should enter the lengths (in morgans) of the chromosome segment of interest (d) and the shared part of this (s) in the second and third rows, respectively (note that s/d is the proportion of shared genome). The programs contain hypothetical values for these and the last row in the output that appears on the screen after the program is executed.

Remark: Our programs evaluate the cumulative probabilities for each s, such that $0 \leq s < d$. Evaluations for the trivial case $s = d$ are not needed because the corresponding cumulative probability is clearly equal to one.

The initial probability vectors are denoted by (c_1, c_2) and (c_1, c_2, c_3, c_4), respectively (*cf.* the last procedure in these programs). These are set up to be the steady-state probabilities. If the user wishes to evaluate cumulative probabilities given information on one of the flanking markers, then he/she should set up the initial probabilities accordingly.

DISCUSSION

In this article we provide Maple V programs for numerical evaluation, with any given accuracy, of the cumulative probabilities of the proportion of genome-shared IBD on chromosome segments by two individuals in either grandchild-grandparent or great-grandchild–great-grandparent relationship. These are the first exact distributional results concerning IBD calculations in the framework of the genomic continuum model. The results also yield exact evaluations of the cumulative probabilities given information (*e.g.*, such as inheritance) on one of the flanking markers. We suggest a couple of applications below. These assume the availability of continuous IBD data. Such data are not yet, but will be made, available with the progress on the Genome Project. In particular, a new technique called genomic mismatch scanning (Nelson *et al.* 1993) is expected to produce almost continuous IBD data (*cf.* Cheung and Nelson 1996; McAllister *et al.* 1996).

Our results can be used in devising tests for identifying chromosomal segments that may contain disease genes. Such segments are expected to have unusually large proportions of genome-shared IBD by the affected related individuals. The cumulative distribution function is the relevant quantitative measure for such unusualness. For example, let a chromosomal segment be suspected of carrying responsible genes for a particular disease. The hypothesis to be tested is "the segment does not carry such genes." Our data consist of observations over the corresponding proportions of genome-shared IBD on that segment for n independent pairs of individuals, each in a grandchild-grandparent relationship, and all affected by the disease. Then a relevant test statistic is the minimum of these proportions. Its cumulative probabilities, and subsequently relevant P values, can be evaluated using the enclosed program for grandchild-grandparent relationship. More specifically, let x be the observed value of this statistic. Then the relevant P value is equal to $(1 - F(x))^n$, where $F(x)$ can be evaluated by the aforementioned program (with $s = xd$, where it should be recalled that d and s are the lengths of the chromosome segment and the shared part, respectively). Note that this test is in the spirit of the QTL tests based on the common theme of allele sharing (*cf.* Lynch and Walsh 1998, pp. 523–533).

Our results can also be used for exact evaluation of P values in significance testing (that is, when there is no specified alternative hypothesis) for resemblance of either a grandchild with a grandparent or a great-grandchild with a great-grandparent. The crossover processes on different chromosomes are assumed to be independent. Our data consist of $x_1, x_2, \ldots, x_n$, where x_i is the observed proportion of genome-shared IBD by the two individuals on the ith autosomal chromosome ($n = 22$ in humans). A relevant test statistic is the maximum of these proportions. Let x be its observed value. Evaluate the corresponding cumulative probabilities (with $s = xd_i$ and d_i the length of the ith autosomal chromosome) for each of these chromosomes. Then the relevant P value is equal to the product of these probabilities. Note that if there is a specified alternative hypothesis, then other tests, based on a full-likelihood approach, would be expected to be more powerful (*cf.* Browning 1998).

Our methodology is general enough to be applicable to higher-order grandparent-type relationships and other relationships. Our methods for deriving explicit closed-form expressions for relevant characteristic functions are applicable to the sojourn time in any given state in any given finite-state Markov chain. Therefore, they are applicable to higher-order grandparent-type relationships. The associated algebra might become unmanageable if the number of states of the underlying Markov chain is too large. However, Poisson approximations (*cf.* Bickeboller and Thompson 1996a,b for the case of half-sibs) should work well in such cases, and therefore exact evaluations might not be needed. Furthermore, similar numerical tools and the help of Maple would yield numerical inversions of such characteristic functions. For other relationships the quantity of interest is the sojourn time in a set of states. This can be identified as a sojourn time in a single state for a suitable Markov renewal process that is associated with the underlying Markov chain. Tools for deriving explicit expressions of the relevant characteristic functions associated with such a sojourn time can be found in Stefanov (1995). We are currently investigating other relationships, such as half-sibs.

The author is grateful to E. Thompson for helpful discussions on the topic and to the referees and the associate editor for their detailed and constructive comments.

LITERATURE CITED

Barndorff-Nielsen, O., 1978 *Information and Exponential Families.* Wiley, Chichester, UK.

Bickeboller, H., and E. A. Thompson, 1996a Distribution of genome shared IBD by half-sibs: approximation by the Poisson clumping heuristic. Theor. Popul. Biol. **50:** 66–90.

Bickeboller, H., and E. A. Thompson, 1996b The probability distribution of the amount of an individual's genome surviving to the following generation. Genetics **143:** 1043–1049.

Brown, L., 1986 *Fundamentals of Statistical Exponential Families.* Institute of Mathematical Statistics, Hayward, CA.

Browning, S., 1998 Relationship information contained in gamete identity by descent data. J. Comp. Biol. **5:** 323–334.

Cheung, V. G., and S. F. Nelson, 1996 Genomic mismatch scanning in human: an identity-by-descent genetic linkage method. Am. J. Hum. Genet. **59** (Suppl.): 1740.

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

Fisher, R. A., 1949 *The Theory of Inbreeding.* Oliver and Boyed, Edinburgh.

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

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

Guo, S. W., 1994b Proportion of genes survived in offspring conditional on inheritance of flanking markers. Genetics **138:** 953–962.

Guo, S. W., 1995 Proportion of genome shared identical by descent by relatives: concept, computation, and applications. Am. J. Hum. Genet. **56:** 1468–1476.

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.

Hill, W. G., 1993 Variation in genetic identity within kinships. Heredity **71:** 652–653.

Lange, K., 1997 *Mathematical and Statistical Methods for Genetic Analysis.* Springer, New York.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

McAllister, L., L. Penland, J. DeRisi and P. O. Brown, 1996 Application of genomic mismatch scanning to mammalian genomes. Am. J. Hum. Genet. **59** (Suppl.): 303.

Monagan, M. B., K. O. Geddes, K. Heal, G. Labahn and S. Vorkoetter, 1997 *Maple V Programming Guide for Release V.* Springer, New York.

Nelson, S. F., J. H. McCusker, M. A. Sander, Y. Kee, P. Modrich *et al.*, 1993 Genomic mismatch scanning: a new approach to genetic linkage mapping. Nat. Genet. **4:** 11–18.

Rice, S. O., 1975 Numerical evaluation of integrals with infinite limits and oscillating integrands. Bell Syst. Tech. J. **54:** 155–164.

Stefanov, V. T., 1991 Noncurved exponential families associated with observations over finite state Markov chains. Scand. J. Stat. **18:** 353–356.

Stefanov, V. T., 1995 Explicit limit results for minimal sufficient statistics and maximum likelihood estimators in some Markov processes: exponential families approach. Ann. Stat. **23:** 1073–1101.

Thompson, E. A., 1995 Genetic importance and genomic descent, pp. 112–123 in *Population Management for Survival and Recovery*, edited by J. D. Ballou, M. Gilpin and T. Foose. Columbia University Press, New York.

Widder, D. V., 1971 *An Introduction to Transform Theory*. Academic Press, New York.

Communicating editor: S. Tavaré

APPENDIX

Proofs of the propositions: The fundamental identity in sequential analysis states that for a finite stopping time the sequential likelihood function is derived from the nonsequential one by substituting the stopping time for the time parameter. Thus, in view of (1), the sequential likelihood function of the chain *X*, observed up to time τ_s, is given by

$$\begin{aligned}\exp\{&N_{12}(\tau_s)\ln(p_1\lambda_1) + N_{13}(\tau_s)\ln(q_1\lambda_1) + N_{21}(\tau_s)\ln(p_2\lambda_2)\\ &+ N_{24}(\tau_s)\ln(q_2\lambda_2) + N_{31}(\tau_s)\ln(p_3\lambda_2)\\ &+ N_{34}(\tau_s)\ln(q_3\lambda_2) + N_{42}(\tau_s)\ln(p_4\lambda_2) + N_{43}(\tau_s)\ln(q_4\lambda_2)\\ &- \lambda_1 T(\tau_s) - \lambda_2 S(\tau_s)\}.\end{aligned}$$

Proof of Proposition 1: Recall that the initial state is 1, that is $X(0) = 1$. Note that the following linear relationships hold:

$$\begin{aligned}N_{12}(\tau_s) + N_{13}(\tau_s) &= N_{21}(\tau_s) + N_{31}(\tau_s)\\ N_{21}(\tau_s) + N_{24}(\tau_s) &= N_{12}(\tau_s) + N_{42}(\tau_s)\\ N_{31}(\tau_s) + N_{34}(\tau_s) &= N_{13}(\tau_s) + N_{43}(\tau_s)\\ T(\tau_s) &= s.\end{aligned}$$

The first three identities compare the number of entries to, with the number of exits from, a state. Note that these numbers are equal if the initial state is 1 (*cf.* Stefanov 1991). For the sake of brevity denote $N_{ij}(\tau_s)$ by N_{ij}. From these identities we can express N_{13}, N_{24}, and N_{34} in terms of the remaining N_{ij}'s, that is,

$$\begin{aligned}N_{13}(\tau_s) &= -N_{12}(\tau_s) + N_{21}(\tau_s) + N_{31}(\tau_s)\\ N_{24}(\tau_s) &= N_{12}(\tau_s) - N_{21}(\tau_s) + N_{42}(\tau_s)\\ N_{34}(\tau_s) &= -N_{12}(\tau_s) + N_{21}(\tau_s) + N_{43}(\tau_s).\end{aligned}$$

Replacing N_{13}, N_{24}, and N_{34} by these we get the following representation of the sequential likelihood function,

$$\exp\left\{\sum_{i=1}^{6}\theta_i Y_i + \phi(\theta)\right\}, \qquad \text{(A1)}$$

where $Y_1 = N_{12}$, $Y_2 = N_{21}$, $Y_3 = N_{31}$, $Y_4 = N_{42}$, $Y_5 = N_{43}$, $Y_6 = S(\tau_s)$, $\theta = (\theta_1, \theta_2, \ldots, \theta_6)$,

$$\begin{aligned}\theta_1 &= \ln\left(\frac{(1-q_1)q_2}{q_1 q_3}\right),\\ \theta_2 &= \ln\left(\frac{q_1(1-q_2)q_3\lambda_1\lambda_2}{q_2}\right),\\ \theta_3 &= \ln(q_1(1-q_3)\lambda_1\lambda_2),\\ \theta_4 &= \ln(q_2(1-q_4)\lambda_2^2),\\ \theta_5 &= \ln(q_3 q_4\lambda_2^2),\\ \theta_6 &= -\lambda_2,\end{aligned}$$

and $\phi(\theta) = -\lambda_1 s$, whose expression in terms of the θ_i's is found below.

In view of Stefanov's (1991) results the family given by (A1) is a noncurved exponential family of order six—that is, there is no linear constraint on the Y_i's and the dimension of the parameter θ is six. For formal definitions and basic analytical properties of exponential families one may refer to Barndorff-Nielsen (1978) or Brown (1986). The characteristic function of the canonical statistic $(Y_1, Y_2, \ldots, Y_6)$ has an explicit representation in terms of $\phi(\theta)$ (*cf.* Barndorff-Nielsen 1978, p. 114). In particular, for the characteristic function of $S(\tau_s)$ we get

$$M_{S(\tau_s)}(t) = \exp(\phi(\theta_1, \theta_2, \ldots, \theta_6) - \phi(\theta_1, \theta_2, \ldots, \theta_5, \theta_6 + It)), \qquad \text{(A2)}$$

where $I = \sqrt{-1}$.

Note that

$$\frac{\theta_2}{\theta_3} = \ln\left(\frac{(1-q_2)q_3}{q_2(1-q_3)}\right).$$

The latter, together with the identities for θ_4 and θ_5 given above, leads to the following expressions for q_2, q_3, and q_4 in terms of the θ_i's:

$$\begin{aligned}q_2 &= \frac{e^{\theta_2-\theta_3+2\theta_4} + e^{\theta_4+\theta_5}}{\theta_6^2 e^{\theta_2-\theta_3+\theta_4} - e^{\theta_2-\theta_3+\theta_4+\theta_5} + e^{\theta_4+\theta_5}},\\ q_3 &= \frac{e^{\theta_2-\theta_3+\theta_4} + e^{\theta_5}}{e^{\theta_2-\theta_3+\theta_4} + \theta_6^2 - e^{\theta_4}},\\ q_4 &= \frac{e^{\theta_2-\theta_3+\theta_4+\theta_5} + \theta_6^2 e^{\theta_5} - e^{\theta_4+\theta_5}}{\theta_6^2(e^{\theta_2-\theta_3+\theta_4} + e^{\theta_5})}.\end{aligned}$$

From the identity for θ_1 we find that $q_1 = q_2/(q_2 + q_3 e^{\theta_1})$; that is,

$$q_1 = \frac{(e^{\theta_2-\theta_3+\theta_4} + \theta_6^2 - e^{\theta_4})(e^{\theta_2-\theta_3+2\theta_4} + e^{\theta_4+\theta_5})}{W_1 + W_2},$$

where

$$\begin{aligned}W_1 &= (e^{\theta_2-\theta_3+\theta_4} + \theta_6^2 - e^{\theta_4})(e^{\theta_2-\theta_3+2\theta_4} + e^{\theta_4+\theta_5}),\\ W_2 &= e^{\theta_1}(e^{\theta_2-\theta_3+\theta_4} + e^{\theta_5})(\theta_6^2 e^{\theta_2-\theta_3+\theta_4} - e^{\theta_2-\theta_3+\theta_4+\theta_5} + e^{\theta_4+\theta_5}).\end{aligned}$$

We denote by $q_i(\theta)$ the expression for q_i in terms of the θ_i's. From the identity for θ_3 we get that

$$\lambda_1 = \frac{e^{\theta_3}}{-\theta_6 q_1(\theta)(1 - q_3(\theta))},$$

and subsequently we derive the following explicit expression for $\phi(\theta)$ $(= -\lambda_1 s)$:

$$\frac{se^{\theta_3}}{\theta_6}\left(\frac{e^{\theta_2-\theta_3+\theta_4} + \theta_6^2 - e^{\theta_4}}{\theta_6^2 - e^{\theta_4} - e^{\theta_5}} + \frac{e^{\theta_1}(e^{\theta_2-\theta_3+\theta_4} + e^{\theta_5})(\theta_6^2 e^{\theta_2-\theta_3+\theta_4} - e^{\theta_2-\theta_3+\theta_4+\theta_5} + e^{\theta_4+\theta_5})}{(\theta_6^2 - e^{\theta_4} - e^{\theta_5})(e^{\theta_2-\theta_3+2\theta_4} + e^{\theta_4+\theta_5})}\right).$$

This leads to an explicit closed-form expression for any particular case. Recall that the case of interest is when $p_i = q_i = 0.5$ for each i and $\lambda_1 = \lambda_2 = 2$. After some easy algebra one gets the following expression for $\phi(\theta_1, \theta_2, \ldots, \theta_5, \theta_6 + It)$:

$$-\frac{2s(1 - It/2)}{2(1 - It/2)^2 - 1}.$$

In view of (A2) this yields Proposition 1.

Proof of Proposition 2: Note that interchanging states two and three do not result in a new transition probability matrix. Therefore, the distribution of $S(\tau_s)$ is the same for both cases $X(0) = 2$ and $X(0) = 3$. Assume the initial state is 2. Similarly to the preceding case we have the linear relationships

$$N_{12}(\tau_s) + N_{13}(\tau_s) = N_{21}(\tau_s) + N_{31}(\tau_s) - 1$$
$$N_{21}(\tau_s) + N_{24}(\tau_s) = N_{12}(\tau_s) + N_{42}(\tau_s) + 1$$
$$N_{31}(\tau_s) + N_{34}(\tau_s) = N_{13}(\tau_s) + N_{43}(\tau_s)$$
$$T(\tau_s) = s.$$

Following similar arguments to those used in the proof of Proposition 1, we derive the same canonical exponential representation as (A1) with a different $\phi(\theta)$. Actually the new $\phi(\theta) = -\lambda_1 s - \ln(q_1 q_3 \lambda_1/q_2)$. Note that if $p_i = q_i = 0.5$ for each i, then

$$\frac{q_1 q_3}{q_2} = \frac{q_1(\theta_1, \ldots, \theta_5, \theta_6 + It)\, q_3(\theta_1, \ldots, \theta_5, \theta_6 + It)}{q_2(\theta_1, \ldots, \theta_5, \theta_6 + It)}.$$

Therefore, if $p_i = q_i = 0.5$ for each i, then the new $\phi(\theta) = -\lambda_1 s - \ln \lambda_1$. Also, the expression for λ_1 in terms of the θ_i's has been found above. Therefore, the proof of proposition 2 is completed similarly to the preceding case.

Proof of Proposition 3: Recall that the initial state is 4, that is $X(0) = 4$. Similarly to the above cases the following identities hold:

$$N_{12}(\tau_s) + N_{13}(\tau_s) = N_{21}(\tau_s) + N_{31}(\tau_s) - 1$$
$$N_{21}(\tau_s) + N_{24}(\tau_s) = N_{12}(\tau_s) + N_{42}(\tau_s)$$
$$N_{31}(\tau_s) + N_{34}(\tau_s) = N_{13}(\tau_s) + N_{43}(\tau_s)$$
$$T(\tau_s) = s.$$

The same canonical representation as (A1), with a different $\phi(\theta)$, is derived. The new $\phi(\theta) = -\lambda_1 s - \ln(q_1 q_3 \lambda_1 \lambda_2)$. Note that if $p_i = q_i = 0.5$ for each i, then

$$\frac{q_1 q_3 \lambda_2}{q_1(\theta_1, \ldots, \theta_5, \theta_6 + It)\, q_3(\theta_1, \ldots, \theta_5, \theta_6 + It)\lambda_2(\theta_1, \ldots, \theta_5, \theta_6 + It)}$$
$$= 1 - \frac{It}{2}.$$

The proof is completed similarly to the above cases.

The proofs of Propositions 4 and 5 follow similar arguments and are therefore omitted. Note also that our methods are based on general results (*cf.* Stefanov 1991) that are valid for any given finite-state Markov chain. Thus, they can also be applied to higher-order grandparent-type relationships.

Details about the inversion of the relevant characteristic functions: First recall the inversion formula

$$F(t_2) - F(t_1) = \lim_{r\to+\infty} \frac{1}{2\pi}\int_{-r}^{r} \frac{e^{-Ist_1} - e^{-Ist_2}}{Is}\psi(s)\,ds,$$

where $\psi(\cdot)$ is the characteristic function of the distribution function $F(\cdot)$ and t_1, t_2 are points of continuity of F. Note that the aforementioned conditional distributions are all supported by the nonnegative part of the real line $[0, +\infty)$. Moreover, the distributions with characteristic functions given in Propositions 2, 3, and 5 are continuous and those with characteristic functions given in Propositions 1 and 4 are mixtures of a continuous part on the interval $(0, +\infty)$ and an atom at zero. Therefore, the relevant inversion formula, for all cases, is

$$F(t) - \frac{p_0}{2} = \lim_{r\to+\infty} \frac{1}{2\pi}\int_{-r}^{r} \frac{1 - e^{-Ist}}{Is}\psi(s)\,ds$$

(*cf.* Theorem 7.1 on p. 106 of Widder 1971), where $p_0 = F(0) - F(0-)$. It is easy to see that

$$p_0 = P(S(\tau_s) = 0 | X(0) = 1) = \exp(-2s)$$

for the conditional distribution of $S(\tau_s)$ given $X(0) = 1$ and

$$p_0 = P(U_2(\nu_s) = 0 | X(0) = 1) = \exp(-s)$$

for the conditional distribution of $U_2(\nu_s)$ given $X(0) = 1$. Therefore, we need to evaluate the integral

$$\int_{-\infty}^{\infty} \frac{e^{-Ist_1} - e^{-Ist_2}}{Is}\psi(s)\,ds$$

for all characteristic functions given in the aforementioned propositions. Some of these can be evaluated directly using the system Maple V as follows. Extract first the real component of the integrand (what one needs to integrate only) and then apply the standard Maple V command for integral evaluation. This is applicable to the characteristic functions $M^{(4)}_{S(\tau_s)}$, $M^{(1)}_{U_2(\nu_s)}$, and $M^{(2)}_{U_2(\nu_s)}$. The integrals associated with $M^{(1)}_{S(\tau_s)}$ and $M^{(2)}_{S(\tau_s)}$ have highly oscillating integrands and cannot be evaluated directly. However, numerical tools, such as those suggested by Rice (1975), turned out to be amenable to these cases. These are based on reducing the oscillation by subtracting from the integrand a suitable integrable function, whose integral is analytically known. The resulting integrands can be evaluated using Maple V. For the evaluation of the integrals associated with $M^{(1)}_{S(\tau_s)}$ and $M^{(2)}_{S(\tau_s)}$ we used $\sin(dx)/x$ and $\cos(dx)/(1 + x^2)$, respectively, where d is the length of the genome of interest. The latter are integrable functions whose integrals are well known:

$$\int_{-\infty}^{\infty} \frac{\sin(dx)}{x} dx = \pi, \quad \int_{-\infty}^{\infty} \frac{\cos(dx)}{1 + x^2} dx = \pi \exp(-d)$$

(*cf.* Rice 1975). Therefore, we can compute the values of the cumulative distribution function of the proportion of genome-shared IBD on chromosome segments by two individuals in either a grandchild-grandparent or a great-grandchild–great-grandparent relationship.

Recall that we assumed $s > 0$ (*cf.* methods). Continuity arguments imply that our formulas also produce the cumulative probability for $s = 0$. An alternative way to show this is by letting τ_0 be the waiting time till the first entry in state 1 and noting that this leads to the same formulas.

Most of the algebra associated with deriving closed-form explicit expressions for the aforementioned characteristic functions could also be done on Maple. This might be necessary in cases of higher-order grandparent-type relationships (Table A1).

TABLE A1

Maple V programs

Grandchild-grandparent relationship

```
> assume (x, real, y, real, c1, real, c2, real);
> d := 3 :
> s := 0.9 :
> expres1 := (1 − exp(−(d − y)*I*x))/I/x*exp(y*I*x/(1 − I*x)) :
> expres1 := simplify(Re(evalc(expres1))) :
> f1 := unapply(expres1, x, y) :
> expres2 := (1 − exp(−(d − y)*I*x))/I/x*exp(y*I*x/(1 − I*x))/(1 − I*x) :
> expres2 := simplify(Re(evalc(expres2))) :
> f2 := unapply(expres2, x, y) :
> for j from 1 to 2 do t · j := evalf(Int(f · j (x, s), x = −infinity . . infinity)) od :
> cumulativeprob := evalf(subs(c1 = 1/2, c2 = 1/2, 1 − ((c1*t1 + c2*t2)/(2*Pi) + c1*exp(−s)/2)));
  cumulativeprob := .2503779941
```

Great-grandchild–great-grandparent relationship

```
> assume (x, real, y, real, c1, real, c2, real, c3, real, c4, real);
> d := 3 :
> s := 2.85 :
> expr1 := (1 − exp(−(d − y)*I*x))/I/x *exp(2*y*(1 − I*x/2)/(2*(1 − I*x/2)* *2 − 1) − 2*y) :
> expr1 := simplify(Re(evalc(expr1))) :
> f1 := unapply(expr1, x, y) :
> g1 := (x, y) − > f1(x, y) − sin((d − y)*x)/x :
> expr2 := (1 − exp(−(d − y)*I*x))/I/x *exp(2*y*(1 − I*x/2)/(2*(1 − I*x/2)* *2 − 1) − 2*y)/(2*(1 − I*x/2)* *2 − 1)*
  (1 − I*x/2) :
> expr2 := simplify(Re(evalc(expr2))) :
> f2 := unapply(expr2, x, y) :
> g2 := (x, y) − > f2(x, y) − cos((d − y)*x)/(1 + x* *2) :
> expr3 := (1 − exp(−(d − y)*I*x))/I/x*exp(2*y*(1 − I*x/2)/(2*(1 − I*x/2)* *2 − 1) − 2*y)/(2*(1 − I*x/2)**
  2 − 1) :
> expr3 := simplify(Re(evalc(expr3))) :
> g3 := unapply(expr3, x, y) :
> for j from 1 to 3 do t · j := evalf(Int(g · j (x, s), x = −infinity . . infinity)) od :
> cumulativeprob := evalf(subs(c1 = 1/4, c2 = 1/4, c3 = 1/4, c4 = 1/4, 1 − ((c1*t1 + (c2 + c3)*t2 + c4*t3)/(2*Pi) +
  (c1 + c1*exp(−2*s) + (c2 + c3)* exp(−d + s))/2))); cumulativeprob := .9980620983
```