BMC Genetics

BioMed Central

Research article

Statistics on continuous IBD data: Exact distribution evaluation for a pair of full(half)-sibs and a pair of a (great-) grandchild with a (great-) grandparent

Valeri T Stefanov

Address: Department of Mathematics and Statistics, The University of Western Australia, Crawley (Perth) 6009, W.A., Australia

E-mail: stefanov@maths.uwa.edu.au

Published: 7 May 2002

BMC Genetics 2002, **3**:7

Received: 21 December 2001
Accepted: 7 May 2002

This article is available from: http://www.biomedcentral.com/1471-2156/3/7

Abstract

Background: Pairs of related individuals are widely used in linkage analysis. Most of the tests for linkage analysis are based on statistics associated with identity by descent (IBD) data. The current biotechnology provides data on very densely packed loci, and therefore, it may provide almost continuous IBD data for pairs of closely related individuals. Therefore, the distribution theory for statistics on continuous IBD data is of interest. In particular, distributional results which allow the evaluation of p-values for relevant tests are of importance.

Results: A technology is provided for numerical evaluation, with any given accuracy, of the cumulative probabilities of some statistics on continuous genome data for pairs of closely related individuals. In the case of a pair of full-sibs, the following statistics are considered: (i) the proportion of genome with 2 (at least 1) haplotypes shared identical-by-descent (IBD) on a chromosomal segment, (ii) the number of distinct pieces (subsegments) of a chromosomal segment, on each of which exactly 2 (at least 1) haplotypes are shared IBD. The natural counterparts of these statistics for the other relationships are also considered. Relevant Maple codes are provided for a rapid evaluation of the cumulative probabilities of such statistics. The genomic continuum model, with Haldane's model for the crossover process, is assumed.

Conclusions: A technology, together with relevant software codes for its automated implementation, are provided for exact evaluation of the distributions of relevant statistics associated with continuous genome data on closely related individuals.

Background

Pairs of related individuals, such as full-sibs, are widely used in linkage analysis. Most of linkage tests are based on statistics associated with identity by descent (IBD) data. Evaluation of p-values requires relevant information on the distributions of such statistics. The current biotechnology provides data on very densely packed loci, and therefore, it may provide almost continuous IBD data for pairs of closely related individuals. The distribution theory for statistics on continuous IBD data has not been developed yet. Bickeboller and Thompson [1,2] provide approximations, based on the Poisson clumping heuristic, to the distribution of the proportion of genome shared IBD by half-sibs, while Stefanov [3] provides a methodology for exact evaluation of the cumulative probabilities for the proportion of genome shared IBD by two individuals in grand-

parent-type relationship. Browning [4,5] suggests a Monte-Carlo approach for such evaluations. Zhao and Liang [6] deal with the exact calculation of the likelihood of a particular relationship for a given gamete IBD data.

This paper provides a technology for numerical evaluation, with any given accuracy, of the cumulative probabilities of relevant statistics for pairs of closely related individuals, such as full(half)-sibs and a (great-)grandchild with a (great-) grandparent. Codes are provided, in the popular software package Maple, for rapidly implementing such evaluations. Possible applications of the results are also discussed (see subsection **Discussion**).

Results and Discussion

Results

A technology is provided for numerical evaluation, with any given accuracy, of cumulative probabilities of statistics on continuous IBD data from pairs of related individuals. The pairs of interest are full-sibs, half-sibs, a grandchild with a grandparent, and a great-grandchild with a great-grandparent. Three Maple codes are provided in the section **Materials and Methods.** Two of these concern a pair of full-sibs. The first one evaluates the cumulative probabilities for the number of pieces, of a chromosomal segment of a fixed length, on each of which 2 haplotypes are shared IBD. The second one evaluates the same for the number of pieces, of a chromosomal segment of a fixed length, on each of which at least 1 haplotype is shared IBD. The third code evaluates the cumulative probabilities for the number of pieces inherited by a great-grandchild from a great-grandparent on a chromosomal segment of a fixed length. The user of these codes should enter the length (in morgans) of the chromosomal segment of interest (y) and the number (k) of pieces. The codes contain hypothetical values for these and the corresponding evaluated probability that appears on the screen after the code is executed. A formula is provided in section **Materials and Methods** (cf. (8)) for a straigthforward evaluation of the cumulative probabilities for the number of IBD pieces for a pair of half-sibs. Also, it is explained in the same section (cf. (1-3) and (6)) how to use the Maple codes provided in [3] in order to evaluate the cumulative probabilities for the proportion of genome with 2 (at least 1) haplotypes shared IBD by a pair of full-sibs, and the proportion of genome shared IBD by a pair of half-sibs, all on a chromosomal segment of a fixed length. Excerpts from such evaluations concerning the statistics of interest for the related pairs of interest are provided in Tables 1,2,3,4,5,6,7,8. Furthermore, our Maple codes evaluate the corresponding cumulative probabilities conditional on information (such as inheritance) on one of the flanking markers. To do so the user should set up the initial probabilities (c_0,c_1,c_2) accordingly.

Table 1: ***Full-sibs:*** **Cumulative probabilities ($F_{T_2(t)/t}(x)$) of the proportion (x) of genome, with 2 haplotypes shared IBD, on a chromosomal segment of length t morgans**

x	$F_{T_2(0.5)/0.5}(x)$	$F_{T_2(1.75)/1.75}(x)$	$F_{T_2(3)/3}(x)$
0.00	0.406268	0.093767	0.021679
0.05	0.452167	0.178482	0.085682
0.10	0.496505	0.271900	0.179515
0.15	0.539142	0.368994	0.294451
0.20	0.579961	0.465228	0.419125
0.25	0.618871	0.556843	0.542321
0.30	0.655800	0.641012	0.655000
0.35	0.690700	0.715865	0.751322
0.40	0.723540	0.780433	0.828748
0.45	0.754309	0.834521	0.887486
0.50	0.783012	0.878543	0.929622
0.55	0.809668	0.913356	0.958215
0.60	0.834310	0.940085	0.976552
0.65	0.856984	0.959985	0.987637
0.70	0.877744	0.974322	0.993929
0.75	0.896655	0.984284	0.997257
0.80	0.913789	0.990929	0.998883
0.85	0.929223	0.995154	0.999603
0.90	0.943039	0.997685	0.999884
0.95	0.955324	0.999085	0.999976
0.99	0.964109	0.999676	0.999997

Discussion

In this article a technology is provided for numerical evaluation, with any given accuracy, of the cumulative probabilities of some statistics on continuous genome data for pairs of related individuals. The pairs of interest are: full-sibs, half-sibs, grandparent with a grandchild, and a great-grandparent with a great-grandchild. In the case of a pair of full-sibs, the following statistics are considered: (i) the proportion of genome with 2 (at least 1) haplotypes shared IBD on a chromosomal segment, (ii) the number of distinct pieces, of a chromosomal segment, on each of which 2 (at least 1) haplotypes are shared IBD. The natural counterparts of these statistics for the other relationships are also covered. Relevant Maple codes are provided for a rapid evaluation of the cumulative probabilities of such statistics.

In the case of full-sibs the IBD is meant within pedigrees consisting of a pair of sibs and their two parents – that is, nuclear families. Also, our distributional results assume such an interpretation of IBD. If the sibs in a pair, or at least one of their parents, are inbred within a larger pedigree than the nuclear family (see [7] for relevant terms) then IBD subsegments with respect to the nuclear family and IBD subsegments with respect to the larger pedigree will not be distinguishable, due to identity-by-state (IBS)

Table 2: ***Full-sibs:*** **Cumulative probabilities ($F_{(T_1(0.5)+T_2(0.5))/0.5}(x)$) of the proportion ($x$) of genome, with at least 1 haplotype shared IBD, on a chromosomal segment of length t morgans**

x	$F_{(T_1(0.5)+T_2(0.5))/0.5}(x)$	$F_{(T_1(1.75)+T_2(1.75))/1.75}(x)$	$F_{(T_1(3)+T_2(3))/3}(x)$
0.00	0.033834	0.000228	0.0000015
0.05	0.044676	0.000915	0.000024
0.10	0.056961	0.002315	0.000116
0.15	0.070777	0.004846	0.000397
0.20	0.086211	0.009071	0.001117
0.25	0.103345	0.015716	0.002743
0.30	0.122256	0.025678	0.006071
0.35	0.143016	0.040015	0.012363
0.40	0.165690	0.059915	0.023448
0.45	0.190332	0.086644	0.041785
0.50	0.216988	0.121457	0.070378
0.55	0.245691	0.165479	0.112514
0.60	0.276460	0.219567	0.171252
0.65	0.309300	0.284135	0.248678
0.70	0.344200	0.358988	0.345000
0.75	0.381129	0.443157	0.457679
0.80	0.420039	0.534772	0.580875
0.85	0.460858	0.631006	0.705549
0.90	0.503495	0.728100	0.820485
0.95	0.547833	0.821518	0.914318
0.99	0.584435	0.890234	0.968161

status of the data. Consequently, if the sibs in a pair, or at least one of their parents, are inbred, then the data will record larger numbers of distinct chromosomal segments, with 2 haplotypes shared IBD, than those in the case of non-inbreeding. Therefore, the distribution of the number of such pieces for a pair of sibs, which is evaluated by the enclosed relevant Maple codes, can be used to assess the evidence for a lack of inbreeding. Such information may be used accordingly. Likewise, the distribution results for the proportion of shared genome and number of pieces IBD on the different chromosomes may be used in testing for mis-specified sib-relationship. Such tests for significance may be based on a combination of separate tests each corresponding to the data on a single chromosome with the suitable Bonferroni correction of the significance level. Similar to these applications hold for a grandparent-type relationship when using the corresponding distribution results.

Our results may also be used in identifying chromosomal segments that may contain loci responsible for complex deseases. Nonparametric tests, similar to that suggested in [3] for pairs in grandparent-type relationship, can be devised for pairs of full-sibs. Assume a chromosomal segment is suspected of carrying responsible gene(s) for a particular desease. The hypothesis to be tested is 'the segment does not carry such genes'. Assume a continuous IBD data are available for n independent pairs of full-sibs, all affected by the desease. In particular, the data contain the proportions, $x_1, x_2,...,x_n$, of genome with 2 haplotypes shared IBD on the chromosomal segment in question. A relevant test statistic is the minimum of these proportions, say x, for these n full-sib pairs. The relevant p-value is equal to $(1 - F(x))^n$, where $F(x)$ can be evaluated using the relevant Maple code. Likewise, a similar test can be based on the corresponding proportions of genome with at least 1 haplotype shared IBD. Both tests are robust and do not depend on the mode of inheritance. However, one may expect that the first one, based on the genome with 2 haplotypes shared IBD, would be more sensitive to a recessive pattern of inheritance on the chromosomal segment. Also, the second one is relevant if sharing of either one or two alleles cannot be distinguished. Our resuts may also be used in identifying the presence of another gene(s) responsible for a complex desease on a chromosomal segment flanked by an already identified major desease gene. The relevant tests are to be based on the corresponding proportions of shared genome, conditional on the information on one of the flanking markers. Recall that our Maple codes also evaluate such conditional probabilities and therefore evaluate the relevant p-values.

Table 3: *Full-sibs:* Cumulative probabilities ($F_{S_2(t)}(k)$) of the number of pieces (k), of a chromosomal segment of length t morgans, on each of which 2 haplotypes are shared IBD

k	$F_{S_2(0.5)}(k)$	$F_{S_2(1.75)}(k)$	$F_{S_2(3)}(k)$
0	0.406268	0.093767	0.021679
1	0.858831	0.364666	0.122635
2	0.985672	0.684247	0.331534
3	0.999253	0.890249	0.587775
4	0.999977	0.972789	0.797866
5	0.9999995	0.995045	0.921127
6	0.999999993	0.999317	0.975280
7	0.9999999999	0.999927	0.993699

Table 4: *Full-sibs:* Cumulative probabilities ($F_{S_{1,2}(t)}(k)$) of the number of pieces (k), of a chromosomal segment of length t morgans, on each of which at least 1 haplotype is shared IBD

k	$F_{S_{1,2}(0.5)}(k)$	$F_{S_{1,2}(1.75)}(k)$	$F_{S_{1,2}(3)}(k)$
0	0.033834	0.000228	0.0000015
1	0.744868	0.187077	0.043355
2	0.972793	0.542254	0.201915
3	0.998551	0.826240	0.461153
4	0.999955	0.954258	0.714397
5	0.9999991	0.991320	0.881336
6	0.999999987	0.998769	0.960918
7	0.9999999999	0.999865	0.989642

Conclusions

A technology, together with relevant software codes for its automated implementation, are provided for exact evaluation of the distributions of relevant statistics associated with continuous genome data on closely related individuals.

Materials and Methods

The underlying mathematical model

Througout the paper the genomic continuum model, with Haldane's model for the crossover process, is assumed. That is, the occurrence of crossovers along the chromosomes is modelled by a Poisson process (see [8]). If the distances are measured in morgans then the rate of the Poisson process is one. Donnelly [9] 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, and time refers to distance. For a pair of full-sibs (the relevant pedigree consists of the two sibs and their parents) the relevant hyperecube is four-dimensional. The coordinates are either 0 or 1 depending on whether a grand-paternal or a grand-maternal DNA was transmitted. The first two coordinates indicate the parental transmissions for sib one and the other two do the same for sib two. For example, the vertex (0,1,1,0) indicates the following transmissions at a chromosomal locus: a grand-paternal (0) from the mother and a grand-maternal (1) from the father of sib one, and a grand-maternal (1) from the mother and a grand-paternal (0) from the father of sib two. The DNA at a location on, or a segment from, one of the homologous chromosomes is called haplotype. The sixteen states of the hypercube can be divided into three groups of vertices indicating whether 0,1, or 2 haplotypes are shared IBD at a locus with the assumption of non-distinguishing between sharing of maternal and paternal DNA. Then the underlying model can be reduced to a three-state continuous time Markov chain whose parameters are described as follows (see [10]). States are denoted by 0,1, and 2 corresponding to the number of shared IBD haplotypes. The holding times are exponentially distributed with rate parameter 4 and the one-step transition probability matrix of the embedded discrete time Markov chain is given by:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{bmatrix}.$$

The initial probability vector is (1/4,1/2,1/4) (the steady-state probabilities). The continuous data on a chromosomal segment consists of the lengths of the consecutive pieces (subsegments) characterised by the number of haplotypes shared IBD.

The sojourn time in a state has the following interpretation. Let d be the length (in morgans) of a chromosome segment of interest. Then the sojourn time in state i (i = 0,1,2), within time interval of length d, is the length of genome whose each location has i haplotypes shared IBD by the two sibs on that segment. Such a genome will be called briefly a genome with i haplotypes shared IBD. The aforementioned sojourn time divided by the length of the segment (d) is the corresponding proportion of genome with i haplotypes shared IBD on the segment. The number of entries to state i, within time interval of length d, is related to the number of distinct pieces (subsegments), of a chromosomal segment of length d, on each of which exactly i haplotypes are shared IBD. The number of entries from state 0 to state 1, within time interval of length d, is related to the number of distinct pieces (subsegments), of a chromosomal segment of length d, on each of which at least 1 haplotype is shared IBD. The latter relationship is explained in the next section.

Table 5: *Half-sibs:* Cumulative probabilities ($F_{T_1(t)/t}(x)$) of the proportion (x) of genome shared IBD on a chromosomal segment of length t morgans

x	$F_{T_1(0.5)/0.5}(x)$	$F_{T_1(1.75)/1.75}(x)$	$F_{T_1(3)/3}(x)$
0.00	0.183940	0.015099	0.001239
0.05	0.212090	0.032818	0.006254
0.10	0.241298	0.057443	0.016951
0.12	0.253256	0.069370	0.023395
0.14	0.265360	0.082525	0.031326
0.16	0.277604	0.096922	0.040906
0.18	0.289981	0.112569	0.052281
0.20	0.302484	0.129460	0.065587
0.22	0.315106	0.147583	0.080936
0.24	0.327840	0.166912	0.098416
0.26	0.340677	0.187415	0.118088
0.28	0.353611	0.209046	0.139980
0.30	0.366634	0.231752	0.164087
0.32	0.379738	0.255469	0.190364
0.34	0.392914	0.280125	0.218734
0.36	0.406157	0.305638	0.249079
0.38	0.419456	0.331920	0.281245
0.40	0.432805	0.358876	0.315045
0.42	0.446194	0.386401	0.350260
0.44	0.459617	0.414390	0.386642
0.46	0.473064	0.442730	0.423920
0.48	0.486528	0.471306	0.461806
0.50	0.500000	0.500000	0.500000

Table 6: *Half-sibs:* Cumulative probabilities ($F_{U(t)}(k)$) of the number of pieces (k) shared IBD on a chromosomal segment of length t morgans

k	$F_{U(0.5)}(k)$	$F_{U(1.75)}(k)$	$F_{U(3)}(k)$
0	0.183940	0.015099	0.001239
1	0.827729	0.228368	0.039660
2	0.988676	0.631039	0.218130
3	0.999661	0.896163	0.525991
4	0.999994	0.981694	0.795609
5	0.99999994	0.997833	0.936728
6	-	0.999818	0.985540
7	-	0.999989	0.997486

Likewise, the underlying model for a pair of half-sibs is a continuous time Markov chain with four states which are the vertices of the two-dimensional cube. The four states can be divided into two groups, each indicating the number (0 or 1) of haplotypes shared IBD at a locus and again not distinguishing between sharing of maternal and paternal DNA. Then the reduced underlying model is a two-state continuous time Markov chain with states denoted by 0 and 1, exponentially distributed holding time with rate parameter 2, and the following one-step transition probability matrix of the embedded discrete-time Markov chain:

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

The initial probability vector is $(1/2, 1/2)$.

Table 7: *Grandchild Grandparent:* Cumulative probabilities ($F_{U(t)}(k)$) of the number of pieces (k) shared IBD on a chromosomal segment of length t morgans

k	$F_{U(0.5)}(k)$	$F_{U(1.75)}(k)$	$F_{U(3)}(k)$
0	0.303265	0.086887	0.024894
1	0.947704	0.610924	0.311169
2	0.999038	0.933144	0.731248
3	0.999992	0.994333	0.941287
4	0.99999997	0.999721	0.992146

Table 8: *Great-grandchild Great-grandparent:* Cumulative probabilities ($F_{S_2(t)}(k)$) of the number of pieces (k) shared IBD on a chromosomal segment of length t morgans

k	$F_{S_2(0.5)}(k)$	$F_{S_2(1.75)}(k)$	$F_{S_2(3)}(k)$
0	0.547465	0.261424	0.125676
1	0.954037	0.692916	0.444327
2	0.998520	0.930617	0.763259
3	0.999978	0.990853	0.932295
4	0.9999998	0.999236	0.986580
5	0.9999999990	0.999957	0.998085

There is a very close similarity between the aforementioned underlying models and those corresponding to the relationships great-grandchild-great-grandparent and grandchild-grandparent. Donnelly (1983) discusses the models for grandparent-type relationships. Note that the reduced underlying model for the relationship great-grandchild-great-grandparent and that for a pair of full-sibs are the same, except for the value of the parameter of the holding time distribution. This is 4 for a pair of full-sibs and 2 for the relationship great-grandchild-great-grandparent. Of course, the interpretation of the states is different. For example, state 2 (or state 0; note that states 0 and 2 are interchangable) indicates a transmission of a great-grandparent DNA to a great-grandchild and the remaining two states indicate two cases resulting in non-transmission. Likewise, the reduced underlying model for

the relationship grandchild-grandparent and that for a pair of half-sibs are the same, again except for the value of the parameter of the holding time distribution. This is 2 for a pair of half-sibs and 1 for the relationship grandchild-grandparent.

Methods

Our methodology is similar to that introduced by Stefanov (2000) who treated grandparent-type relationships. Namely, relevant stopping times are introduced and explicit expressions for their characteristic functions are found. These characteristic functions are numerically invertable using the system Maple V (for introduction to Maple V see [11]) and some numerical tools. Therefore, their distribution functions are derivable. Finally, the latter distribution functions yield the distribution functions of relevant random quantities, such as the sojourn time in a state, counts of transitions from a state to another state, all within a fixed time interval. Subsequently, the cumulative probabilities of relevant statistics on continuous IBD data can be calculated. For example, such statistics for a pair of full-sibs are the proportion of genome with 2 (at least 1) haplotypes shared IBD, and the count of pieces, of a chromosomal segment, on each of which 2 (at least 1) haplotypes are shared IBD. More details follow.

Full-sibs

Let $\{X(t)\}_{t\geq 0}$ be a three-state continuous time Markov chain whose parameters are those of the underlying model for a pair of full-sibs. Denote by $N_{ij}(t)$ the number of one-step transitions from state i to state j, and by $T_i(t)$ the sojourn time in state i, i, j = 0,1, 2, all up to time t.

Note the following interpretation of these quantities when considering a chromosomal segment of length t. The sojourn time in state 2, $T_2(t)$, is the length of genome with both haplotypes shared IBD, on a chromosomal segment of length t; the sojourn time $T_1(t) + T_2(t)$ is the length of genome with at least one haplotype shared IBD, on a chromosomal segment of length t; $N_{12}(t)$ ($N_{12}(t)$ + 1), given the initial state is not 2 (is 2), counts the number of distinct pieces, on each of which both haplotypes are shared IBD, on a chromosomal segment of length t; $N_{01}(t)$ ($N_{01}(t)$ +1), given the initial state is 0 (is not 0), counts the number of distinct pieces, on each of which at least one haplotype is shared IBD, on a chromosomal segment of length t.

Note the following interpretation of $T_2(t)$ and $N_{01}(t)$ if the aforementioned three-state Markov chain is the underlying model for the great-grandchild-great-grandparent relationship: $T_2(t)$ is the amount of genome inherited by a great-grandchild from his great-grandparent on a chromosomal segment of length t, and $N_{12}(t)$ ($N_{12}(t)$ +1), given the initial state is not 2 (is 2), is the number of distinct pieces whose relevant haplotypes are shared IBD on a chromosomal segment of length t.

Denote by $F_t(s)$ the cumulative probability that is evaluated by the Maple program provided in [3] for the great-grandchild-great-grandparent relationship (t and s are the lengths of the chromosomal segment and the shared IBD part of it, respectively). Then it is easy to see that the following hold when the underlying model is that for a pair of full-sibs:

$$P\left(T_2(t)\leq x\right)=F_{2t}(2x), \tag{1}$$

$$P\left(T_1(t)+T_2(t)\leq x\right)=1-F_{2t}\left(2(t-x)\right), \quad \text{if } x>0, \tag{2}$$

$$P\left(T_1(t)+T_2(t)=0\right)=\frac{\exp(-4t)}{4}. \tag{3}$$

Therefore, in view of the aforementioned interpretation of $T_2(t)$ and $T_1(t) + T_2(t)$, using the identities (1), (2), and (3), and the Maple program for the great-grandchild-great-grandparent relationship provided in [3], one can derive the cumulative probabilities of the following quantities associated with a pair of full-sibs: the proportion of genome with both haplotypes shared IBD, and the proportion of genome with at least one haplotype shared IBD, on any chromosomal segment.

In what follows we discuss how the cumulative probabilities of other relevant statistics are derived. Introduce the following stopping times:

$$\tau_k=\inf\left\{t: N_{12}(t)=k\right\}, \qquad k\geq 1, \quad \tau_0=0,$$

$$\upsilon_k=\inf\left\{t: N_{01}(t)=k\right\}, \qquad k\geq 1, \quad \upsilon_0=0.$$

Explicit expressions for the characteristic functions of τ_k and ν_k corresponding to different initial states are derivable. The relevant propositions and their proofs are found in the next subsection.

Denote by $S_2(t)$ ($S_{1,2}(t)$) the number of distinct pieces of a chromosomal segment of length t, on each of which 2 (at least 1) haplotypes are shared IBD. Then the distributions of $S_2(t)$ and $S_{1,2}(t)$ are related to the distributions of the τ_k and ν_k as follows.

$$P\left(S_2(t)\leq k\right)=1-\sum_{i=0}^{1}P\left(\tau_{k+1}\leq t \mid X(0)=i\right)P\left(X(0)=i\right) - P\left(\tau_k\leq t \mid X(0)=2\right)P\left(X(0)=2\right), \quad k=0,1,\ldots. \tag{4}$$

$$P\left(S_{1,2}(t)\leq k\right)=1-\sum_{i=1}^{2}P\left(\upsilon_k\leq t \mid X(0)=i\right)P\left(X(0)=i\right) - P\left(\upsilon_{k+1}\leq t \mid X(0)=0\right)P\left(X(0)=0\right), \quad k=0,1,\ldots. \tag{5}$$

In order to compute these cumulative probabilities we need the conditional cumulative probabilities of the τ_k

and ν_k given the initial state. The propositions in the next subsection provide the characteristic functions of these conditional distributions. They are numerically invertable, and subsequently, the required cumulative probabilities are derivable. Likewise, the cumulative probabilities of the number of distinct pieces inherited by a great-grandchild from a great-grandparent on a chromosomal segment of length t can be calculated (see Remark 2 in the next subsection).

The relevant Maple codes, for rapidly implementing such evaluations, are provided in subsection **Maple V codes.**

Half-sibs

Consider now the underlying model for a pair of half-sibs. We use the same notation, $N_{ij}(t)$ and $T_i(t)$, (i,j = 0,1), for the number of one-step transitions from state i to state j and the sojourn time at state i, respectively, up to time t. The sojourn time $T_1(t)$ is the amount of genome shared IBD by the half-sibs on a chromosomal segment of length t. Similarly to the preceding case the cumulative probabilities of the proportion of such genome can be calculated using the identity

$$P(T_1(t) \le x) = P(2T_1(t) \le 2x) = F_{2t}(2x), \qquad (6)$$

where $F_t(s)$ is the cumulative probability that is evaluated by the Maple program provided in [3] for the grandchild-grandparent relationship (t and s are the lengths of the chromosomal segment and the shared IBD part of it, respectively).

Introduce the following stopping times

$$\mu_k = \inf\{t : N_{01}(t) = k\}, \quad k = 1,2,..., \quad \mu_0 = 0,$$

that is, μ_k is the waiting time till entering state 1 for the k-th time. Denote by $U(t)$ the number of IBD pieces on a chromosomal segment of length t. Then it is easy to see that the following hold.

$$P(U(t) \le k \mid X(0) = 1) = 1 - P(\mu_k \le t \mid X(0) = 1), \quad k = 0,1,...,$$
$$P(U(t) \le k \mid X(0) = 0) = 1 - P(\mu_{k+1} \le t \mid X(0) = 0), \quad k = 0,1,...$$

Therefore, the distribution of $U(t)$ is related to the distributions of the μ_k, as follows:

$$P(U(t) \le k) = \sum_{i=0}^{1} \left(1 - P(\mu_{k-i+1} \le t \mid X(0) = i)\right) P(X(0) = i). \qquad (7)$$

It is easy to see that μ_k, given the initial state is 1, is distributed as the sum of 2k independent and exponentially distributed random variables with parameter 2. Likewise μ_k, given the initial state is 0, is distributed as the sum of $2k - 1$ such variables. Therefore, the following hold.

Fact 1. The conditional distribution of μ_k, given the initial state is 1, is a Gamma distribution ($G(2k, 0.5)$) with parameters $2k$ and 0.5.

Fact 2. The conditional distribution of μ_k, given the initial state is 0, is a Gamma distribution ($G(2k - 1,0.5)$) with parameters $2k - 1$ and 0.5.

In view of these facts and the identity given in (7)

$$P(U(t) \le k) = 1 - c_0 F_{G(2k+1,0.5)}(t) - c_1 F_{G(2k,0.5)}(t), \qquad (8)$$

where $F_{G(.,.)}$ is the cumulative distribution function of a Gamma distribution G(.,.) and (c_0,c_1) is the initial probability vector. Thus, the cumulative probabilities of $U(t)$ can be computed using any standard statistical software. Excerpts of such probabilities are provided in Table 6 in the case $c_0 = c_1 = 0.5$ (the steady-state probabilities).

Remark 1. If the underlying model for the grandchild-grandparent relationship is considered then the second parameter of the aforementioned Gamma distributions is to be changed from 0.5 to 1.

Relevant characteristic functions

Consider the underlying model for a pair of full-sibs. The random quantities τ_k and ν_k have been introduced in subsection **Methods.** The following propositions hold.

Proposition 1. Assume that the initial state is either 0 or 2. Then the characteristic function of τ_k is given by:

$$M_{\tau_k}^{(0)}(s) = \frac{1}{\left(\left(2\left(1 - \frac{Is}{4}\right)^2 - 1\right)\right)^k}, \qquad k = 1,2...,$$

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

Proposition 2. Assume that the initial state is 1. Then the characteristic function of τ_k is given by:

$$M_{\tau_k}^{(1)}(s) = \frac{1 - \frac{Is}{4}}{\left(\left(2\left(1 - \frac{Is}{4}\right)^2 - 1\right)\right)^k}, \qquad k = 1,2....$$

Proposition 3. Assume that the initial state is 0. Then the characteristic function of ν_k is given by:

$$M_{\upsilon_k}^{(0)}(s) = \frac{1}{\left(1-\frac{Is}{4}\right)\left(\left(2\left(1-\frac{Is}{4}\right)^2-1\right)\right)^{k-1}}, \qquad k = 1, 2....$$

Proposition 4. Assume that the initial state is 1. Then the characteristic function of ν_k is given by:

$$M_{\upsilon_k}^{(1)}(s) = \frac{1}{\left(\left(2\left(1-\frac{Is}{4}\right)^2-1\right)\right)^{k}}, \qquad k = 1, 2....$$

Proposition 5. Assume that the initial state is 2. Then the characteristic function of ν_k is given by:

$$M_{\upsilon_k}^{(2)}(s) = \frac{1}{\left(1-\frac{Is}{4}\right)\left(\left(2\left(1-\frac{Is}{4}\right)^2-1\right)\right)^{k}}, \qquad k = 1, 2....$$

Let the assumptions of Proposition 1 be satisfied. Note that if the chain starts from state 0 then after two transitions it will be either in state 2 or back in state 0. Each of these two outcomes has probability 0.5. Therefore, the distribution of τ_1 is a geometric sum of independent and identically distributed random variables whose distribution is Gamma with parameters 2 and 1/4. The support of the geometric distribution is {1,2,...} and its parameter equals 0.5. Likewise, all of the above apply if the chain starts from state 2. The characteristic function of the above Gamma distribution is given by $(1 - Is/4)^{-2}$, and the generating function of the geometric distribution is given by (s/2)/(1 - s/2). From well-known results on generating functions of random sums (see [12]) the characteristic function of τ_1 is given by the composition of the latter two, that is

$$M_{\tau_1}(s) = \frac{1}{\left(2\left(1-\frac{Is}{4}\right)^2-1\right)}.$$

Note that the process regenerates at the stopping times τ_k. Thus, the characteristic function of τ_k is the k-th power of the above expression. This completes the proof of Proposition 1.

The statement of Proposition 2 (recall that the initial state is 1) follows from the fact that in this case τ_k is shorter than its counterpart from Proposition 1 by an independent and exponentially distributed random variable with parameter 4.

The proofs of the remaining propositions follow similar arguments and are therefore omitted.

Remark 2. If the underlying model for the great-grandchild-great-grandparent relationship is considered then in the above expressions one should change

$$\frac{Is}{4} \text{ by } \frac{Is}{2}.$$

Also $S_2(t)$ means now the number of distinct pieces, inherited by a great-grandchild from his great-grandparent, on a chromosomal segment of length *t*.

Maple V codes

Full-sibs

Cumulative probabilities for the number of pieces (k), of a chromosomal segment of length γ morgans, on each of which 2 haplotypes are shared IBD

```
> assume(x, real, γ, real):

> γ := 0.5:

> k := 2 :

> ex1 := (1 - exp(-γ * I * x))/I/x/((2 * (1 - I * x/4) * *2 - 1)
* *(k + 1)) :

> ex1 := simplify(Re(evalc(ex1))) :

> f1 := unapply(ex1, x) :

> ex2 := (1 - exp(-γ * I * x))/I/x * (1 - I * x/4)/((2 * (1 - I *
x/4) * *2 - 1) * *(k + 1)):

> ex2 := simplify(Re(evalc(ex2))) :

> f2 := unapply(ex2,x) :

> ex3 := (1 - exp(-γ * I * x))/I/x/((2 * (1 - I * x/4) * *2 - 1)
* *k) :

> ex3 := simplify(Re(evalc(ex3))) :

> f3 := unapply(ex3, x) :
```

```
> for j from 1 to 3 do a.j := eval f(Int(f.j(x),x = -infinity..infinity)) od :

> c0 := 1/4 :

> c1 := 1/2 :

> c2 := 1/4 :

> if k = 0 then eval f(1 - (a1 * c0 + a2 * c1)/(2 * Pi) - c2)
elseeval f(1 - (a1 * c0 + a2 * c1 + a3 * c2)/(2 * Pi)) fi;

.9856720668
```

Full-sibs

Cumulative probabilities for the number of pieces (k), of a chromosomal segment of length γ morgans, on each of which at least 1 haplotype is shared IBD

```
> assume(x, real, γ, real):

> γ := 0.5:

> k := 1 :

> ex1 := (1 - exp(-γ * I * x))/I/x/((1- I * x/4) * (2 * (1 - I *
x/4) * *2 - 1) * *k) :

> ex1 := simplify(Re(evalc(ex1))) :

> f1 := unapply(ex1, x) :

> ex2 := (1 - exp(-γ * I * x))/I/x/((2 * (1 - I * x/4) * *2 - 1)
* *k):

> ex2 := simplify(Re(evalc(ex2))) :

> f2 := unapply(ex2,x) :

> for j from 1 to 2 do a.j := eval f(Int(f.j(x),x = -infinity..infinity)) od :

> c0 := 1/4 :

> c1 := 1/2 :

> c2 := 1/4 :

> if k = 0 then evalf(1 - (a1 * c0)/(2 * Pi) - c1 - c2) else
evalf(1 - (a1 * (c0 + c2) + a2 * c1)/(2 * Pi)) fi;

.7448682221
```

Great-grandchild-great-grandparent

Cumulative probabilities for the number of pieces (k), inherited by a great-grandchild from his great-grandparent, on a chromosomal segment of length γ morgans

```
> assume(x, real, γ, real):

> γ := 0.5:

> k := 3 :

> ex1 := (1 - exp(-γ * I * x))/I/x/((2 * (1 - I * x/2) * *2 - 1)
* *(k + 1)) :

> ex1 := simplify(Re(evalc(ex1))) :

> f1 := unapply(ex1, x) :

> ex2 := (1 - exp(-γ * I * x))/I/x * (1 - I * x/2)/((2 * (1 - I *
x/2) * *2 - 1) * *(k + 1)):

> ex2 := simplify(Re(evalc(ex2))) :

> f2 := unapply(ex2,x) :

> ex3 := (1 - exp(-γ * I * x))/I/x/((2 * (1 - I * x/2) * *2 - 1)
* *k) :

> ex3 := simplify(Re(evalc(ex3))) :

> f3 := unapply(ex3, x) :

> for j from 1 to 3 do a.j := eval f(Int(f.j(x),x = -infinity..infinity)) od :

> c0 := 1/4 :

> c1 := 1/2 :

> c2 := 1/4 :

> if k = 0 then eval f(1 - (a1 * c0 + a2 * c1)/(2 * Pi) - c2)
elseeval f(1 - (a1 * c0 + a2 * c1 + a3 * c2)/(2 * Pi)) fi;

.9999781100
```

Acknowledgements

This research is supported by a grant from UWA under research grant number RA/1/485/10. The author is very grateful to E. Thompson for helpful discussions on the topic.

References

1. Bickeboller H, Thompson EA: **Distribution of genome shared IBD by half-sibs: Approximation by the Poisson clumping heuristic.** *Theor. Popul. Biol.* 1996, **50:**66-90
2. Bickeboller H, Thompson EA: **The probability distribution of the amount of an individual's genome surviving to the following generation.** *Genetics* 1996, **143:**1043-1049

3. Stefanov VT: **Distribution of genome shared identical by descent by two individuals in grandparent-type relationship.** *Genetics* 2000, **156:**1403-1410
4. Browning S: **Relationship information contained in gamete identity by descent data.** *J. Comp. Biol.* 1998, **5:**323-334
5. Browning S: **A Monte Carlo approach to calculating probabilities for continuous identity by descent data.** *J. Appl. Prob.* 2000, **37:**850-864
6. Zhao H, Liang F: **On relationship inference using gamete identity by descent data.** *J. Comput. Biol.* 2000, **8:**191-200
7. Thompson EA: **Pedigree analysis in human genetics.** *Baltimore, Johns Hopkins University Press* 1986
8. Lange K: **Mathematical and statistical methods for genetic analysis.** *New York, Springer* 1997
9. Donnelly K: **The probability that related individuals share some section of the genome identical by descent.** *Theor. Popul. Biol.* 1983, **23:**34-64
10. Ethier SN, Hodge SE: **Identity-by-descent analysis of sibship configurations.** *Amer. J. Med. Genet.* 1985, **22:**263-272
11. Monagan MB, Geddes KO, Heal K, Labahn G, Vorkoetter S: **Maple V programming guide for release V.** *New York, Springer* 1997
12. Grimmett GR, Stirzaker DR: **Probability and random processes.** *Second Edition, Oxford, Clarendon Press* 1994