Available online at www.sciencedirect.com

ScienceDirect

Theoretical Population Biology 73 (2008) 264–276

Theoretical Population Biology

www.elsevier.com/locate/tpb

Recombination, gene conversion, and identity-by-descent at three loci

Danielle Jones*, John Wakeley

Department of Organismic and Evolutionary Biology, Harvard University, 4092-4100 Biological Laboratories, 16 Divinity Avenue, Cambridge, MA, 02138, United States

Received 10 July 2007
Available online 4 December 2007

Abstract

We investigate the probabilities of identity-by-descent at three loci in order to find a signature which differentiates between the two types of crossing over events: recombination and gene conversion. We use a Markov chain to model coalescence, recombination, gene conversion and mutation in a sample of size two. Using numerical analysis, we calculate the total probability of identity-by-descent at the three loci, and partition these probabilities based on a partial ordering of coalescent events at the three loci. We use these results to compute the probabilities of four different patterns of conditional identity and non-identity at the three loci under recombination and gene conversion. Although recombination and gene conversion do make different predictions, the differences are not likely to be useful in distinguishing between them using three locus patterns between pairs of DNA sequences. This implies that measures of genetic identity in larger samples will be needed to distinguish between gene conversion and recombination.

Keywords: Identity-by-descent; Multi-locus identity-by-descent; Coalescent; Gene conversion; Recombination

1. Introduction

In sexual organisms meiotic crossing over is an important device for evolution; crossing over shuffles existing genomic associations to produce potentially superior novel genotypic combinations. The differential resolution of the Holliday junction (Stahl, 1994), created during prophase I of meiosis, produces two types of crossing over events: gene conversion and recombination. We adopt the definition of Wiuf and Hein (2000): if the resolution leads to the reciprocal exchange of flanking regions along the chromosome, then it is recombination; if the resolution leads to the non-reciprocal replacement of a homologous region without concomitant exchange of flanking regions, then it is gene conversion.

It is currently of great interest to distinguish between gene conversion and recombination, and to measure the rates of each along a genome (Padhukasahasram et al., 2004; Przeworski and Wall, 2001; Ardlie et al., 2001; Song et al., 2006). The present lack of such a method limits our ability to determine the marker density necessary to find disease-associated loci using linkage disequilibrium (Padhukasahasram et al., 2004; Ardlie et al., 2001; Hernández-Sánchez et al., 2006; Hein et al., 2005). In addition, without such a method it is impossible to make useful comparisons of gene conversion and recombination within and between individual genomes, populations (Frisse et al., 2001), and species or higher taxa (Jorde, 2005). Here we investigate a new way to discriminate between these two types of crossover events based on polymorphism at three loci. Specifically, we consider patterns of Identity-By-Descent (IBD) at three loci in a sample of two chromosomes. We use the definition of Malecot (1975): alleles are IBD if they descend from a common ancestor without any intervening mutations.

Determining the rate of fine scale crossing over in breeding experiments is difficult due to the rarity of the initiation and completion of crossover events, especially over such short distances as the tract length of a gene conversion event. As Pritchard and Przeworski (2001) point out, the traditional approach of analyzing pedigrees greatly limits the number of meioses, and thus the possible recombination events, that can be examined. To accurately measure crossing over requires the examination of an enormous number of

* Corresponding author.
E-mail addresses: djones@fas.harvard.edu (D. Jones), wakeley@fas.harvard.edu (J. Wakeley).

0040-5809/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2007.10.006

meioses. The most feasible approach is sperm-typing, which further restricts the measurements to crossing over in males. A comprehensive study of male crossover rates involved genotyping individual sperm at a large panel of markers in the Major Histocompatibility Complex (MHC) (Jeffreys et al., 2001). Jeffreys and May (2004) subsequently analyzed single-nucleotide polymorphisms (SNPs) located in known hotspots for recombination, such as the MHC and PAR1 region, to measure the ratio of gene conversion to recombination events. They concluded that 80%–94% of all crossing over events resolve themselves as gene conversions. The length of the non-reciprocal replacement is referred to as the tract length of the gene conversion event. In humans, the tract length has been estimated to be between about 50 and 500 base pairs (bp) (Jeffreys and May, 2004; Padhukasahasram et al., 2004). Tract lengths may differ between species but appear consistent within individual species (Jeffreys and May, 2004).

Population genetics, in contrast to the experimental techniques mentioned above, provides an alternative framework in which to obtain a large number of meiosis. With the tools of population genetics it is possible to sample distantly related individuals. This strategy increases the length of time and so the number of meiotic events surveyed. It also has the benefit of comparing the inferred rate of gene conversion in both sexes, under different demographic parameters and on a genomic scale.

Linkage Disequilibrium (LD) measures the association, within a population, between alleles at different locations along a chromosome. Both gene conversion and recombination act to break up these associations, and thus decrease LD, but in different ways depending on the distance between loci. Polymorphic markers that are far apart (suggested by Andolfatto and Nordborg (1998) to be > a tract length apart, see Eq. (1) of their paper) will be predominantly affected by recombination events, because gene conversion events whose entire tracts are between the markers have no effect on LD between the markers. LD between markers that are close together (<1 kb) will be affected by both gene conversion and recombination. Thus, LD is expected to decay sharply at short distances, but more slowly as the distance between markers increases. A number of studies have observed patterns consistent with this (Frisse et al., 2001; Ardlie et al., 2001; Przeworski and Wall, 2001; Andolfatto and Nordborg, 1998).

It should then be possible to distinguish the effects of gene conversion and recombination using pairs of loci, and most studies that use polymorphism data to measure gene conversion take a pairwise approach (Ptak et al., 2004; Frisse et al., 2001; Andolfatto and Nordborg, 1998; Ardlie et al., 2001). One such method, based on a pairwise composite likelihood approach, was developed by Hudson (2001). This method calculates the likelihood of parameters of haplotype data sets by calculating the likelihoods of SNP pairs and then multiplying these values between all possible pairs. Frisse et al. (2001) and Ptak et al. (2004) used this approach to estimate the ratio of gene conversion to recombination in humans (assuming a tract length of 500 bp) to be between 1 (Ptak et al., 2004) and 7.3 (Frisse et al., 2001).

However, because a single gene conversion event has the same effect as a double recombination event, we expect that three loci will be preferable to two loci in distinguishing gene conversion from recombination. Padhukasahasram et al. (2004) recently developed a simulation-based maximum likelihood method using summary statistics at two and three loci to estimate the ratio of gene conversion to recombination. The summary statistics were carefully chosen to reflect the two processes of crossing over, and the method was applied to data from human chromosome 21. The ratio of gene conversion to recombination was estimated to be 9.4 (Padhukasahasram et al., 2004).

The analysis we undertake here is in same vein as the computational work of Padhukasahasram et al. (2004). We show that some patterns of three-locus identity are enriched under gene conversion compared to recombination. This follows from the construction and numerical analysis of a Markov chain describing the ancestry of three loci in a sample of two chromosomes in the presence of gene conversion, recombination, and mutation. In particular, we focus on patterns of identity that are most likely when the two outer loci have a shared ancestry that is decoupled from the ancestry at the middle locus.

A few papers have recently presented methods to calculate IBD at multiple loci in the presence of recombination and mutation. Hill and Weir (2007) leverage the relationship between non-IBD and LD at multiple loci to reduce the complexity of the transition probabilities of sampling and recombination involved in the calculation of IBD at multiple loci over time. In order to avoid the complexities encountered at more than three loci (Hill and Weir, 2007), Hill and Hernández-Sánchez (2007) develop an approximate method to predict the non-IBD at multiple loci based on the chain rule which, in turn, extends the regression model of Hernández-Sánchez et al. (2006). The method we use here is similar to that of Hill and Weir (2007), and of the earlier work of Strobeck and Morgan (1978). The difference is that we include gene conversion in addition to recombination.

In a model with three loci, where gene conversion at the middle locus might theoretically be differentiated from recombination based on patterns of IBD, the size of the tract length will impact the ability to distinguish between gene conversion and recombination. The ideal situation would seem to be that each locus is much smaller than the tract length while the distances between the loci are greater than the tract length. Then, most gene conversion events will either convert an entire locus or have no effect at all because both ends of the tract sit between two of the loci. For simplicity, we do not include a tract length as part of our model, but instead assume that the effect of gene conversion is to convert single loci. A depiction of our model is given in Fig. 1. Gene conversion of the outer loci (A or C) has the same effect as a single recombination event, while conversion of the middle locus (B) is indistinguishable from a double recombination event.

The rest of the paper is as follows. We define a Markov chain to describe the movement of chromosomes between ancestral states (see Fig. 2 for an example of one time step),

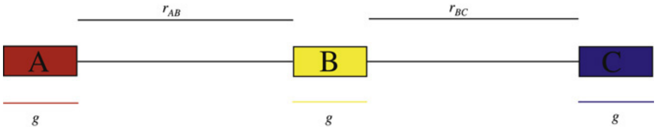


Fig. 1. Depiction of the three-locus model. The two types of crossing over events are restricted to specific parts of the considered area: gene conversion can only occur at the loci themselves, where it occurs at equal rates, and recombination can only occur in the two intervening spaces between the loci. Recombination is additive across the considered area: $r_{AC} = 1-(1-r_{AB})(1-r_{BC})$. Mutation, which is neutral, can only occur at each of the three loci and it occurs with the same rate at each locus.

ultimately leading to coalescence at all three loci. We then include mutation in a single-generation transition matrix, **P**, and its coalescent, time-rescaled counterpart, **Q**. We can compute probabilities of IBD using either **P** or **Q**. We combine the three-locus results with results for one and two loci in order to compute patterns of identity and non-identity that we expect will distinguish between gene conversion and recombination. We plot these quantities over a range of parameter values, and also compare our numerical results with the results of simulations using the ms (make sample) program of Hudson (2002). As expected there is an enrichment of particular states under conditions of gene conversion versus recombination. Although the differences are not dramatic, the results are useful in delineating cases in which we expect three-locus patterns to be useful and in suggesting directions for future work.

2. Materials & methods

2.1. The model and probabilities of IBD

Forward in time, Wright–Fisher reproduction is a two step process. In the first step, a diploid individual produces haploid gametes. During the production of these gametes, the opportunity for mutation, recombination, and gene conversion exists. The gametes subsequently contribute to a population wide gamete pool. In the second step, the next generation is formed by sampling $2N$ sequences with replacement from the gamete pool. Recombination and gene conversion act to place neighbouring regions from separate ancestors, which possess potentially different histories, onto the same chromosome. In the model presented here recombination events are restricted to the two long interlocus regions; gene conversion events, as well as mutation events, are limited to each of the three loci (see Fig. 1). Mutations occur according to an infinite alleles model (*i.e.* no back mutation), so that a mutated locus is not IBD. We assume that all variation is selectively neutral.

We follow the ancestry of a sample of two chromosomes which each possess three loci: A, B, and C. Each generation backwards in time consists of two steps which have the opposite order of their forward-time counterparts: the first is the possibility of coalescence, and the second is the possibility of a crossing over event (shown in Fig. 2). Backwards in time, coalescent events place sampled chromosomes onto the same ancestor whereas crossing over events separate the

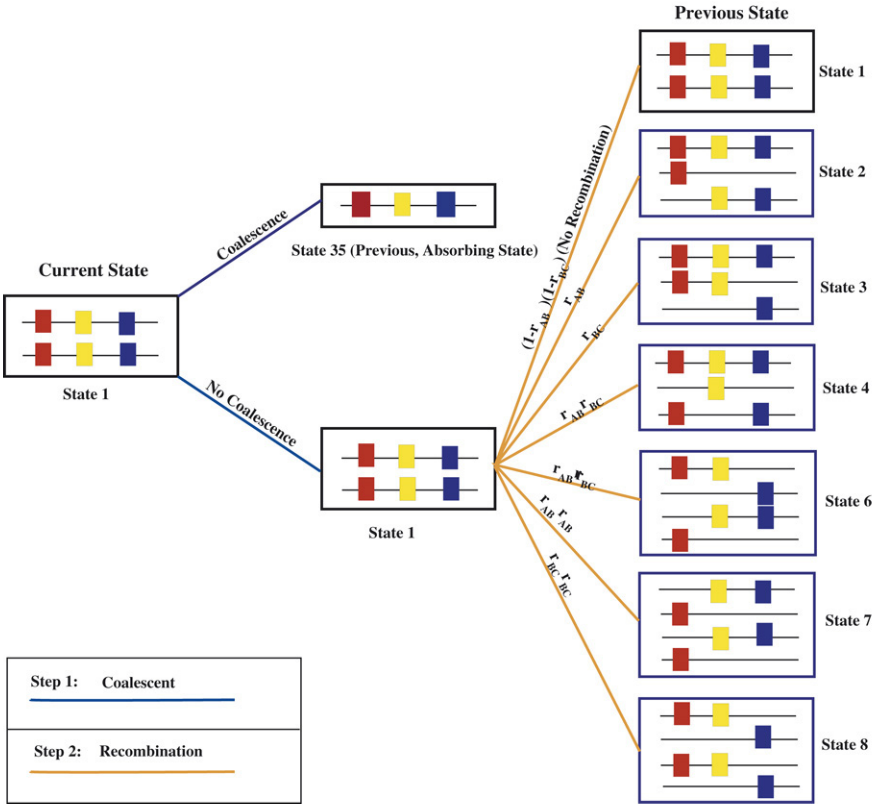
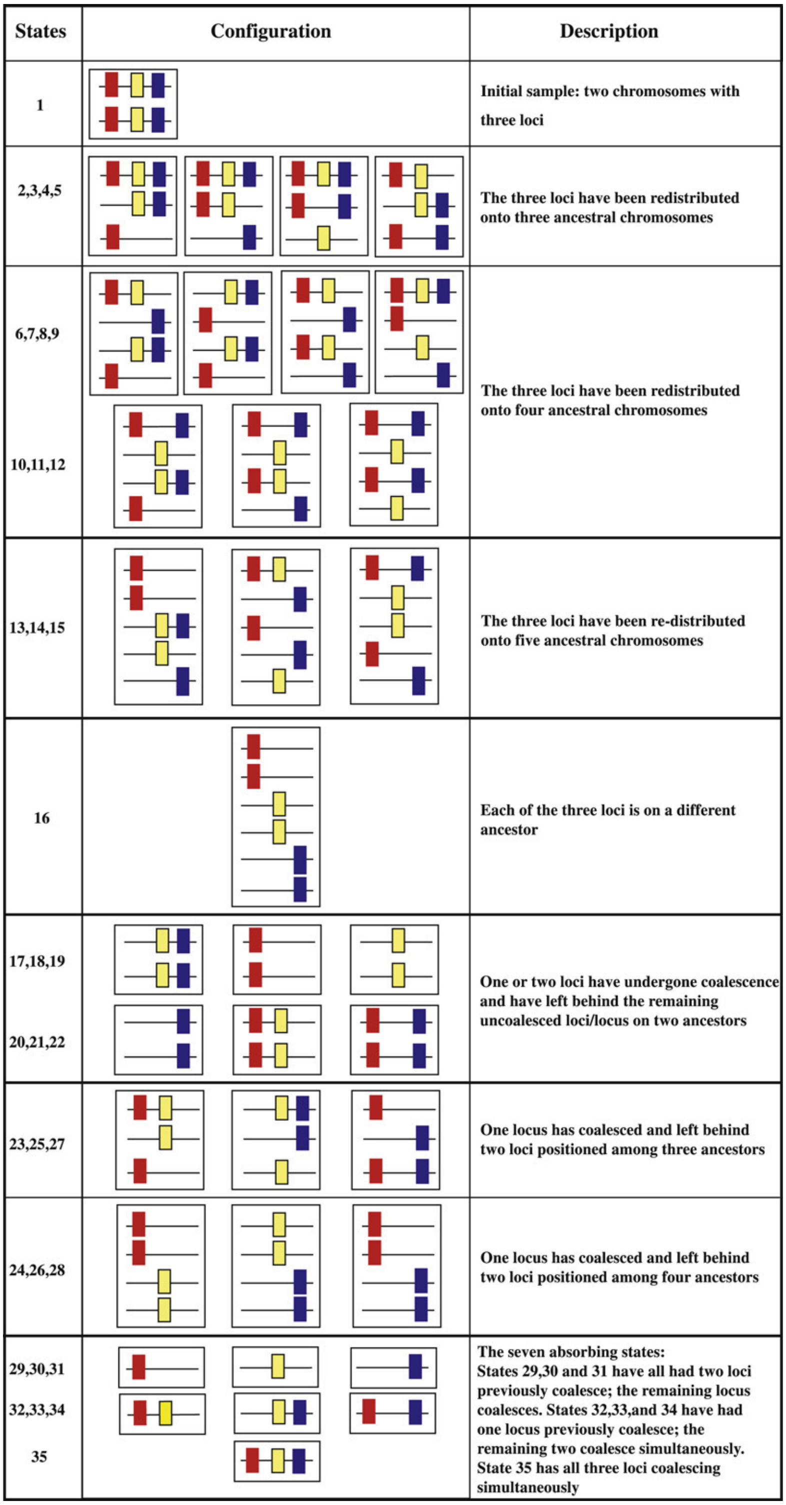


Fig. 2. The events of one time step with no mutation and no gene conversion. In one time step back in time, the sample can either coalesce or, if it doesn't coalesce, it can undergo recombination (or not) to produce its ancestral states.



States	Configuration	Description
1		Initial sample: two chromosomes with three loci
2,3,4,5		The three loci have been redistributed onto three ancestral chromosomes
6,7,8,9 10,11,12		The three loci have been redistributed onto four ancestral chromosomes
13,14,15		The three loci have been re-distributed onto five ancestral chromosomes
16		Each of the three loci is on a different ancestor
17,18,19 20,21,22		One or two loci have undergone coalescence and have left behind the remaining uncoalesced loci/locus on two ancestors
23,25,27		One locus has coalesced and left behind two loci positioned among three ancestors
24,26,28		One locus has coalesced and left behind two loci positioned among four ancestors
29,30,31 32,33,34 35		The seven absorbing states: States 29,30 and 31 have all had two loci previously coalesce; the remaining locus coalesces. States 32,33,and 34 have had one locus previously coalesce; the remaining two coalesce simultaneously. State 35 has all three loci coalescing simultaneously

Fig. 3. The state space for the three-locus model. Whenever a locus coalesces, it leaves the other non-coalescing and non-mutating loci (represented by shaded boxes) behind on ancestors. This convention is not followed in the case of the seven absorbing states where the remaining loci are shown as shaded boxes despite being in the process of coalescing.

loci and places them on different ancestors. For simplicity only recombination is shown in Fig. 2, but this step could involve either gene conversion instead of recombination or both recombination and gene conversion. Mutation is also included in the model but is not shown in Fig. 2.

We chose to keep track of 35 possible states of the sample, and these are shown in Fig. 3. Constructing the Markov chain in this way allowed us to partition the probability of IBD at all three loci into seven different components, corresponding to seven absorbing states which differ in the order of coalescent events that have occurred (Fig. 3: states 29 through 35). We began by writing down the matrix $\mathbf{P}$ of single-generation transition probabilities for the ancestry of the sample without mutation.

The entries of $\mathbf{P}$ are calculated as follows. First, a pair of chromosomes has a probability of $1/(2N)$ to coalesce. Next, there is the possibility of either recombination or gene conversion. The probability of a recombination event between locus A and locus B is r_{AB}, and the probability of a recombination event between locus B and C is r_{BC}. Since a recombination that occurs between locus A and locus C must occur in either of these two segments, we do not use an additional symbol r_{AC}; implicitly, $r_{AC} = 1 - (1 - r_{AB})(1 - r_{BC})$. The probabilities of gene conversion are g_A, g_B, and g_C. For a visual representation of how the transition probabilities were calculated, including coalescence and recombination, see Fig. 2. We do not display $\mathbf{P}$, or any of the other 35×35 matrices below, but a Mathematica (Wolfarm Research, Inc., 2002) file containing them is available from the authors upon request.

Of the 35 states in the chain, seven are absorbing (Fig. 3, states 29 through 35). When one of these seven states is entered, all three loci have reached their common ancestors, but in a different order depending on which of the seven states is entered. State 35 occurs when all three loci coalesce in the same generation. The six other states involve at least one of the loci finding its common ancestor before the last locus coalesces. For example, state 31 represents the event that locus C coalesces and that locus A and locus B have already coalesced. State 32 is the event that locus A and locus B simultaneously coalesce and locus C has already coalesced.

The entries in each row of $\mathbf{P}$ sum to one. This exact transition matrix includes the probabilities of some events that are unlikely when the population size is large and the probabilities of recombination and gene conversion are small. Thus, we also constructed a rate matrix $\mathbf{Q}$ for the continuous-time version of the model. In particular, we imagine that N is very large, and rescale time so that it is measured in units of $2N$ generations, as typical in coalescent models (Nordborg, 2001). Thus, we define the rate matrix by the limit

$$\mathbf{Q} = \lim_{N \to \infty} [2N(\mathbf{P} - \mathbf{I})], \tag{1}$$

assuming that

$$2Nr_{AB} \to \frac{\rho_{AB}}{2}, \qquad 2Nr_{BC} \to \frac{\rho_{BC}}{2}, \qquad 2Ng_A \to \frac{\kappa}{2}, \tag{2}$$

as N tends to infinity. We assume that $g_A = g_B = g_C$, so there is only one gene conversion rate, κ. The entries in each row of $\mathbf{Q}$ sum to zero. When N is large and the probabilities of recombination and gene conversion are small, we expect the τ-generation transition matrix $\mathbf{P}^{\tau}$ to be well approximated by the matrix exponential $\mathrm{e}^{t\mathbf{Q}}$, where $\tau = 2Nt$. All of the calculations presented below were computed in both ways to verify the accuracy of this approximation.

To include mutation in the model, we constructed a vector whose 35 entries were the probabilities of not mutating for each state. There are only four possible values for the entries in this mutation-probability vector: $(1 - v)^6$ for states that have three uncoalesced loci, $(1 - v)^4$ for states that have two uncoalesced loci, $(1 - v)^2$ for states that only have one uncoalesced locus, and one for each of the seven absorbing states. A new matrix, $\mathbf{P}_{\text{mut}}$, was obtained by multiplying the entries in each row of $\mathbf{P}$ by the entry in the corresponding row of the mutation-probability vector. When mutation is included in this way, the entries in each row of the resulting matrix $\mathbf{P}_{\text{mut}}$ no longer add to one. We can imagine an eighth absorbing state (not represented in the matrix) which is the event that at least one locus is not IBD. Finally, we obtained a coalescent rate matrix $\mathbf{Q}_{\text{mut}}$ from $\mathbf{P}_{\text{mut}}$ as described above, with the additional assumption that $2N\mu \to \theta/2$ as N tends to infinity.

As one example of the entries in these matrices consider the transition between state 1 and state 2 (Fig. 3):

$$\begin{aligned}(\mathbf{P}_{\text{mut}})_{12} = {} & 2(1 - 1/(2N))(1 - u)^6(r_{AB}(1 - r_{AB}) \\ & \times (1 - r_{BC})^2(1 - g_A)^2(1 - g_B)^2(1 - g_C)^2 \\ & + (1 - r_{AB})^2(1 - r_{BC})^2 g_A \\ & \times (1 - g_A)(1 - g_B)^2(1 - g_C)^2).\end{aligned}$$

It is possible to move from state 1 to state 2 with one recombination event between locus A and locus B (r_{AB}) or one gene conversion event at locus A (g_A). In comparison, a second example is:

$$\begin{aligned}(\mathbf{P}_{\text{mut}})_{14} = {} & 2(1 - 1/(2N))(1 - u)^6(r_{AB}r_{BC}(1 - r_{AB}) \\ & \times (1 - r_{BC})(1 - g_A)^2(1 - g_B)^2(1 - g_C)^2 \\ & + (1 - r_{AB})^2(1 - r_{BC})^2 g_B(1 - g_A)^2 \\ & \times (1 - g_B)(1 - g_C)^2).\end{aligned}$$

In this case, the move from state 1 to state 4 requires two recombination events, one between locus A and locus B (r_{AB}) and one between locus B and locus C (r_{BC}) on the same chromosome, or one gene conversion event at locus B (g_B). The corresponding, coalescence-rescaled rates become $(\mathbf{Q}_{\text{mut}})_{12} = \rho_{AB} + \kappa_A$, which includes one recombination event and one gene conversion event, and $(\mathbf{Q}_{\text{mut}})_{14} = \kappa_B$, which only includes a possible gene conversion event at locus B since the simultaneous recombination events have a negligible probability.

We are interested in the equilibrium probabilities of IBD for the three loci, for a sample in state one, *i.e.* two chromosomes with three loci. These are obtained in the limit as t tends to infinity, and are the last seven entries in first row of the matrix $\lim_{t \to \infty} \mathbf{P}^t$ or $\lim_{t \to \infty} \mathrm{e}^{t\mathbf{Q}}$. Again, the last seven states, 29 through 35, are absorbing, and are reached when all three loci have coalesced. When the possibility of mutation is included, as in $\mathbf{P}_{\text{mut}}$ and $\mathbf{Q}_{\text{mut}}$, states 29 through 35 are reached only when no mutations have occurred at any of the three loci. These seven probabilities can be obtained by solving the equation $\mathbf{x} = \mathbf{x}\mathbf{P}$, in which case they are the last seven entries in the vector $\mathbf{x}$. General analytical solutions to any of these equations are unavailable, so all of the calculations below are numerical. The sum of the equilibrium probabilities of being in each of the seven absorbing states is equal to the probability of IBD. We compared these seven probabilities over a range of values of the parameters of the model.

In some of the calculations below, we require the corresponding two-locus probabilities of IBD. Since the two-locus case is contained in our 35 × 35 matrices, we could obtain these probabilities from our analysis. However, Strobeck and Morgan (1978) gives the closed-form solution for this probability in their Eq. (5). We used their equation, substituting in the appropriate values for recombination. For example, for the probability of IBD between locus A and locus C with recombination, but no gene conversion, the recombination parameter is $\rho_{AB}+\rho_{BC}$; while if gene conversion is included, it would be $2\kappa+\rho_{AB}+\rho_{BC}$. We also checked that we could obtain Strobeck and Morgan (1978) results from a reduced version of our matrices.

2.2. Distinguishing recombination from gene conversion

We considered the ability of four different conditional probabilities of patterns of identity and non-identity among the three loci to distinguish between gene conversion and recombination. All four quantities could be computed from haplotype data. The first two were designed to capture the extent to which the two outer loci might have a different history than the middle locus. These measures should pick up the differential effects of gene conversion on the middle locus (*i.e.* like two recombination events) and on the two outer loci (*i.e.* like simple recombination). The second two quantities compared the probabilities at locus C to probabilities at locus A and locus B, with increasing distance between locus B and locus C. The four probabilities are shown below; their derivation can be found in the Appendix.

We use E_A, E_B, and E_C to denote the events whereby locus A, locus B or locus C are IBD. The complements of these events, denoted using a superscript 'c', are the respective probabilities of non-identity at each locus. The two measures that contrast ancestries at the middle locus and at the outer loci are as follows. We considered the probability that the middle locus is not IBD given that the two outer loci are IBD. This is given by:

$$P(E_B^c|E_A,E_C)=\frac{P(E_A,E_C)-P(E_A,E_B,E_C)}{P(E_A,E_C)}. \qquad (3)$$

Next, we considered the probability that the middle locus is IBD, given that the two outer loci are not IBD. This is given by:

$$P(E_B|E_A^c,E_C^c)=\frac{P(E_B)-P(E_B,E_C)-P(E_A,E_B)+P(E_A,E_B,E_C)}{1-P(E_A)-P(E_C)+P(E_A,E_C)}. \qquad (4)$$

The derivations of these quantities are straightforward, but are relegated to the Appendix.

The second two quantities are the probability that locus C is IBD given that locus A and locus B are not IBD,

$$P(E_C|E_A^c,E_B^c)=\frac{P(E_C)-P(E_B,E_C)-P(E_A,E_C)+P(E_A,E_B,E_C)}{1-P(E_B)-P(E_A)+P(E_A,E_B)}, \qquad (5)$$

and the probability that locus C is not IBD given that locus A and locus B are IBD,

$$P(E_C^c|E_A,E_B)=\frac{P(E_A,E_B)-P(E_A,E_B,E_C)}{P(E_A,E_B)}. \qquad (6)$$

The idea behind these last two probabilities is to illustrate the differential effects of increasing the distance between loci, specifically how far locus C is from locus A and locus B which are assumed to be close together. If gene conversion is the only factor, these two conditional probabilities will not depend on distance, but they will depend on distance when recombination can occur.

2.3. Simulations using Hudson's program

We checked for consistency between our new numerical results for three loci and the results of simulations using the well-known coalescent program ms (Hudson, 2002). Because the model of gene conversion in ms is different than ours, we restricted our simulations to the case of recombination only. We ran ms with sample of size two and a locus comprised of three sites, with a recombination rate of $\rho_{AB}+\rho_{BC}$ (in fact $\rho_{AB}=\rho_{BC}$ in all of our simulations), and a mutation rate of 3θ.

The output of the program was screened and the numbers of every possible pattern of IBD at the three loci were counted. Letting 0 represent IBD and 1 represent non-IBD, the counts of the eight possible patterns are n_{000}, n_{001}, n_{010}, n_{011}, n_{100}, n_{101}, n_{110} and n_{111}. For example, n_{101} is the number of simulation replicates in which locus B was IBD, but locus A and locus C were not IBD. We used these counts to calculate the values of $P(E_B|E_A^c,E_C^c)$ and $P(E_B^c|E_A,E_C)$ in the simulations. Specifically, $P(E_B|E_A^c,E_C^c)$ is given by $n_{101}/(n_{101}+n_{111})$ and $P(E_B^c|E_A,E_C)$ is given by $n_{010}/(n_{010}+n_{000})$. In all cases, these simulations agreed well with our numerical analysis (see below).

3. Results

Under a broad range of parameter values, we examined the values of the statistics above (Eq. (3) through (6)) and the probability of IBD broken down into its seven constituents (Fig. 3: states 29 through 35).

3.1. Constituents of the equilibrium probability of IBD

At equilibrium, the probability of IBD at all three loci is composed of the probabilities of being in each of the seven states, 29 through 35, of the Markov chain. As expected, the overall probability of IBD decreases with increasing mutation rate, θ. This is shown in Fig. 4, which plots these seven probabilities for the recombination and mutation parameters shown in Table 1. As discussed above, a mutation at any locus negates the possibility of IBD at all three loci. The values of ρ_{AB} and ρ_{BC} also have an effect on the overall probability of IBD, but the effect is minor compared to the effect of θ. The rates of gene conversion were set to zero in all cases in Fig. 4.

Fig. 4 shows that the values of ρ_{AB} and ρ_{BC} have a substantial effect on the relative values of the seven constituents

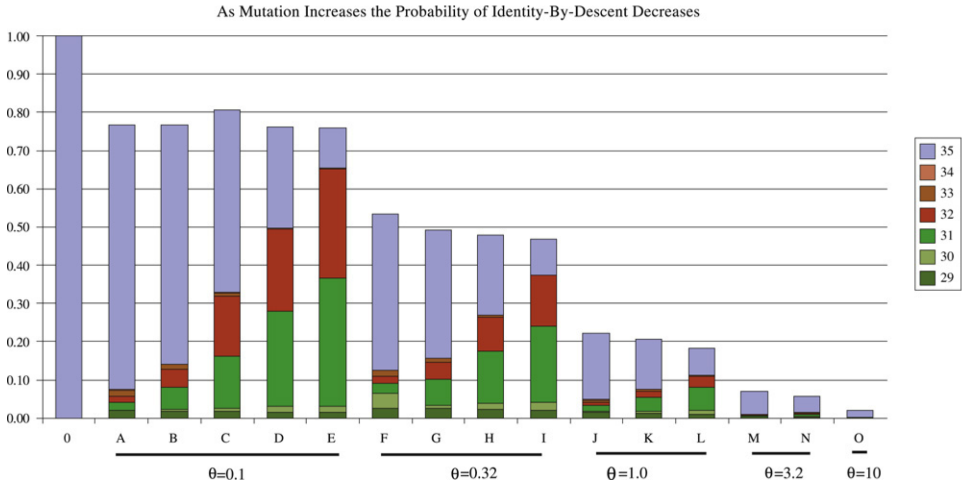


Fig. 4. The effects of mutation rate on the overall equilibrium probability of IBD. As mutation increases, the number of ancestors who have not undergone a mutational event dwindles. Letters *A* through *O* correspond to recombination and θ values found in Table 1. The number 0 indicates a zero value for all parameters. Bars for each state are given from top to bottom in the same order as in the legend of the figure.

Table 1
The 15 different scenarios for recombination and mutation used in producing Fig. 4

Scenario	ρ_{AB}	ρ_{BC}	θ
A	0.1	0.1	0.1
B	0.1	0.32	0.1
C	0.1	1.0	0.1
D	0.1	3.2	0.1
E	0.1	10.	0.1
F	0.32	0.32	0.32
G	0.32	1.0	0.32
H	0.32	3.2	0.32
I	0.32	10.	0.32
J	1.0	1.0	1.0
K	1.0	3.2	1.0
L	1.0	10.0	1.0
M	3.2	3.2	3.2
N	3.2	10.0	3.2
O	10.0	10.0	10.0

of the probability of IBD. Generally speaking, when the rates of recombination are low, the bulk of the probability of IBD is contained in the probability of reaching state 35 (without mutation). When the rates of recombination are higher, there is a greater chance of reaching other states. These more subtle effects are important when we consider the differential effects of gene conversion versus recombination.

Fig. 5 shows the partitioning of the probability of IBD for a single mutation rate ($\theta = 0.1$) as a function of the rate of gene conversion (absent recombination) or recombination (absent gene conversion). The bars from left to right on each side of the figure are comparable in the sense that each has the same rate of crossover between loci, so $\kappa = 0.01$ (on the left) is equivalent to $\rho = 0.02$ (on the right). Overall, the two sides of the graph are similar in their transition from 100% state 35 to roughly equal proportions of states 29, 30, and 31. However, the bars in the middle range of crossover rates ($\kappa \approx 0.1$–3.2, or $\rho_{AB} = \rho_{BC} \approx 0.2$–6.4) offer some hope of distinguishing gene conversion from recombination.

In particular, states 34 and 30 have higher probabilities under gene conversion versus recombination. As can be seen from Fig. 2, state 34 involves the simultaneous coalescence of the two outer loci, the middle locus having coalesced previously. We expect state 34 to be enriched by gene conversion since it requires a restrictive series of events under recombination only. As noted by Wiuf and Hein (2000), it requires two recombination events, one between locus *A* and *B* and a second between locus *B* and *C*, followed by a coalescent event between the two particular ancestors that carry only locus *A* and one which contains only locus *C*. This is much less probable than a single gene conversion event, which achieves the same result. In contrast, state 30 can result from a much looser order of events; it is produced from one gene conversion event acting on a pair of chromosomes in state 1 (in which case it is the complement of state 34) or it can be produced from one recombination event on chromosomes in which one of the outer loci has already coalesced or recombined in a previous generation. The numerous ways of achieving state 30 without gene conversion suggest that state 30 is a poor choice for differentiating gene conversion and recombination. Therefore, our focus is on state 34 in much of what follows.

The relationship between increasing crossing over events and an increase in state 34 is not monotonic. When the rate of conversion (or recombination) becomes much larger than the coalescent rate, such that all of the loci are placed on distinct ancestors, then configuration 34 becomes unlikely. Fig. 6 shows that, for different values of θ, the maximum fraction of the total probability of IBD made up of state 34, and thus the greatest differentiation between the recombination and gene conversion scenarios, occurs when the rate of gene conversion (or recombination) is approximately equal to the rate of coalescence. For the case of $\theta = 0.01$ and $\theta = 0.1$, this results in approximately 50 times the amount of state 34 than is present in the recombination only case, under equivalent conditions. When $\theta = 1.0$, this ratio is 31.

Since mutation, crossing over and coalescence are competing events in this model, it is expected that when θ is high the

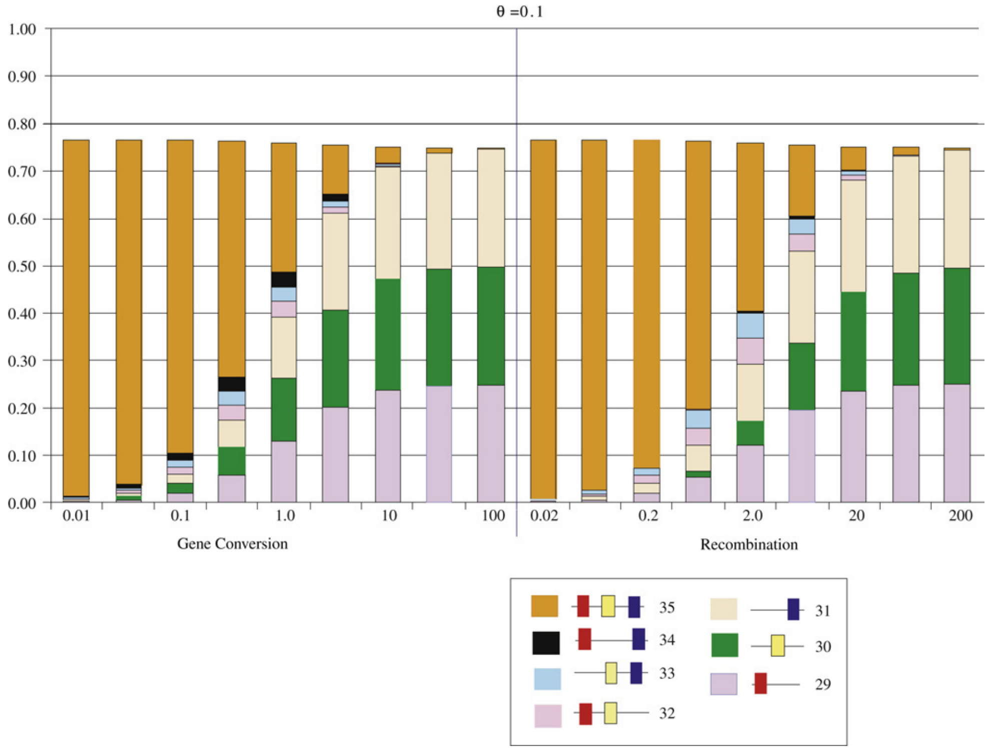


Fig. 5. The effects of two scenarios, recombination with no gene conversion and an equivalent amount of gene conversion with no recombination, on the equilibrium probability of IBD. Bars for each state are given from top to bottom in the same order as in the legend of the figure.

probability of IBD will be composed mainly of state 35. If the loci do not coalesce immediately, before they have a chance to separate onto different ancestors, then they risk mutation. This is the situation seen in Fig. 6 when $\theta = 100$. When $\theta = 10$, the probability of IBD is composed of state 35 until the crossing over rate overtakes the rate of coalescence and is approximately equal to the mutation rate. At this point, for gene conversion, all seven of the states are present in the probability of IBD. As the rate of crossing over overtakes the mutation rate, the loci which do not coalesce immediately (in state 35) and do not mutate are found on three separate ancestors.

We can use the estimates $N = 10^4$, $u = 1.29 \times 10^{-8}$ and $r = 2.58 \times 10^{-8}$ per base pair per generation (Frisse et al., 2001; Jeffreys and May, 2004; Innan et al., 2003), to relate our parameters to numbers of base pairs in humans. Specifically, $\theta = 0.01, 0.1, 1.0, 10$, and 100 correspond to about 20, 200, 2000, 20 000, and 200 000 base pairs, respectively; and $\rho = 0.01, 0.1, 1.0, 10$, and 100 correspond to about 10, 100, 1000, 10 000, and 100 000 base pairs, respectively. Since, as discussed above, we implicitly assume that each locus is smaller than, and that the loci are farther apart than, the typical gene conversion tract, we expect our results for $\theta \leqslant 0.1$ and $\rho_{AB} \geqslant 2.0$ to be roughly applicable to human genetic data.

3.2. *Comparing the probability of different patterns of history at the three loci*

Gene conversion acts on each locus whereas recombination acts between the loci. As a result, it is easier to decouple the histories of the middle locus and the two outer loci with gene conversion than with recombination. The conditional probabilities of four different patterns of identity-by-descent and non-identity-by-descent, two of which (Eqs. (3) and (4)) involve a different history at the middle locus than at the outer two loci, are examined under two distance criteria.

In the first case (shown in Fig. 7) $P(E_B|E_A^c, E_C^c)$ and $P(E_B^c|E_A, E_C)$ are calculated under a range of values for gene conversion, recombination and mutation. The distances between locus A and locus B and between locus B and locus C are assumed to be the same. These probabilities are compared to those produced by `ms` simulations, with recombination (Hudson, 2002). In the second case (shown in Fig. 8) two types of identity-by-descent and non-identity-by-descent probabilities are examined as the distance between locus B and locus C increases: the probability of a different history at the two outer loci and the middle locus and the probability of a different history at locus C and two neighbouring loci (locus A and locus B). In both cases the crossing over rates are scaled to have the same value so, for example, when $\kappa = 0.1$ then $\rho = 0.2$.

A mutated locus can be thought of as possessing a long branch in comparison to the short branch of an unmutated (or IBD) locus. Branch lengths between linked loci are correlated but become less so as crossing over rates increase. In two of the four probabilities examined, $P(E_B|E_A^c, E_C^c)$ and $P(E_B^c|E_A, E_C)$, the middle locus will have on average a different branch length than the two outer loci. Gene conversion is able to separate the history of a middle locus from the histories of the two outer loci with one event. In contrast, to achieve the same state with recombination requires a complex series of events.

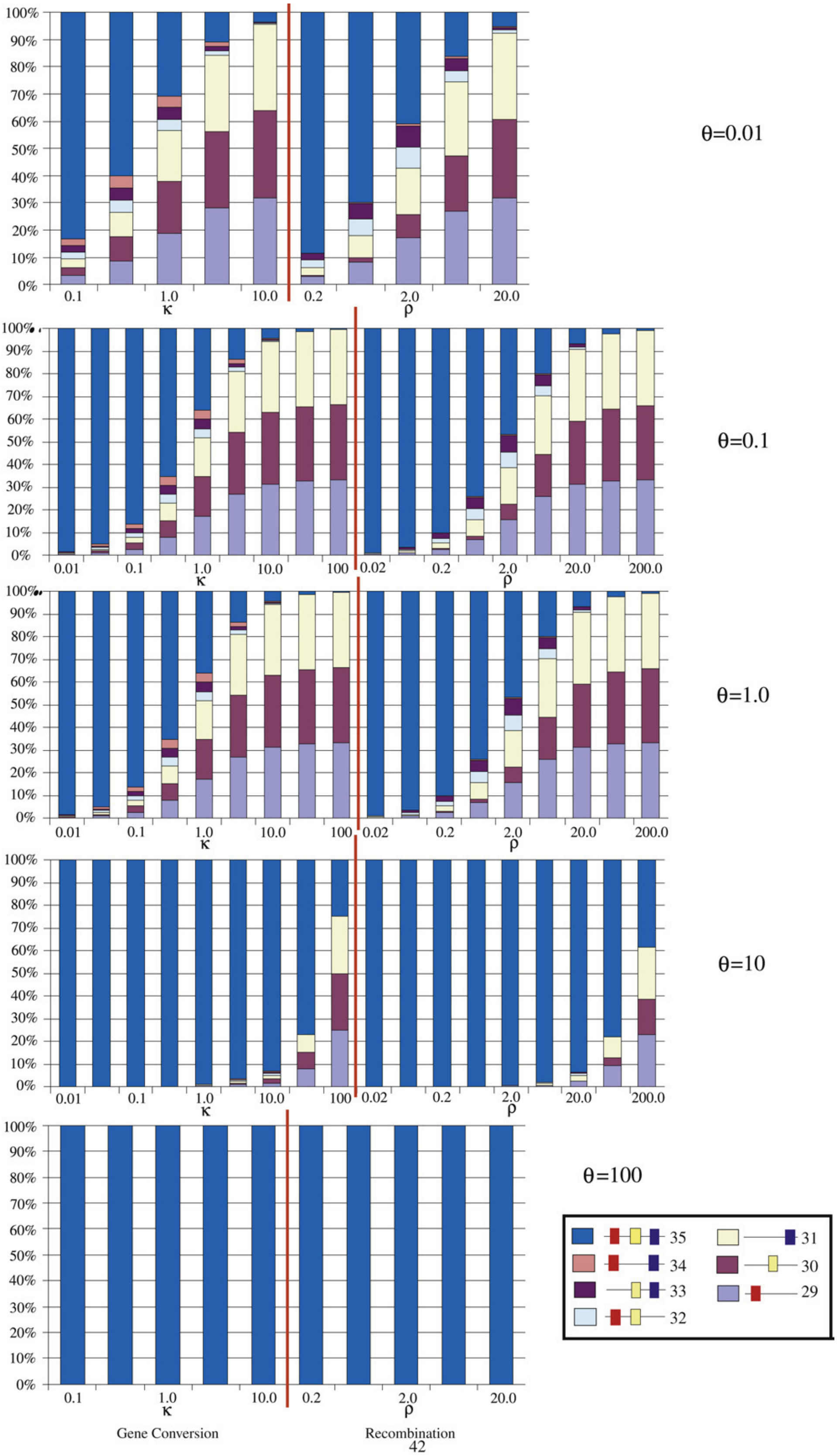


Fig. 6. The effect of increasing mutation rate on the components of the total probability of IBD. These graphs have been rescaled to 100% to increase the visibility of the components of the total probability. The rate of coalescence is 1. Bars for each state are given from top to bottom in the same order as in the legend of the figure.

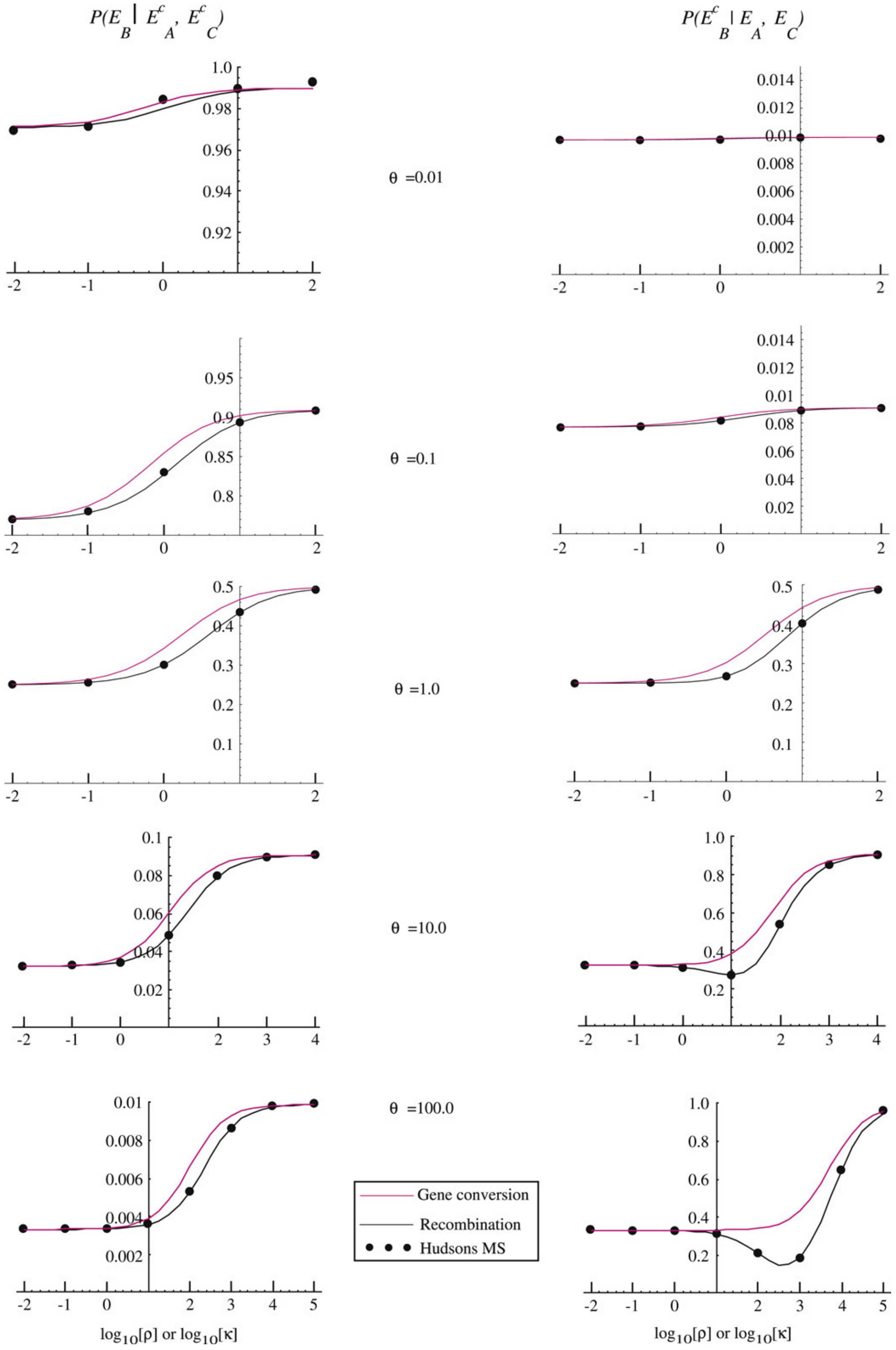


Fig. 7. When the distance between locus A and B and between locus B and C are equal and additive three scenarios are explored under two probability questions: Gene conversion with no recombination using the matrix derived in this paper — thin line, Recombination with no gene conversion using the matrix derived in this paper — thick line, probabilities found using Hudson's ms program — dotted points. The X-axis is the rate of gene conversion in terms of exponents of 10. Therefore, '-2' represents $10^{-2} = 0.01$. This is multiplied by '2' for recombination.

From Fig. 7 it is clear from the different vertical axes of each graph that mutation rate has, in general, an opposite effect on $P(E_B|E^c_A, E^c_C)$ and $P(E^c_B|E_A, E_C)$; as θ increases, for any given value of crossing over, $P(E_B|E^c_A, E^c_C)$ decreases whereas $P(E^c_B|E_A, E_C)$ increases. These two different behaviours result from the conditional nature of the examined probabilities. For instance, $P(E_B|E^c_A, E^c_C)$ is conditioned on a mutation occurring at both locus A and locus C so as θ increases,

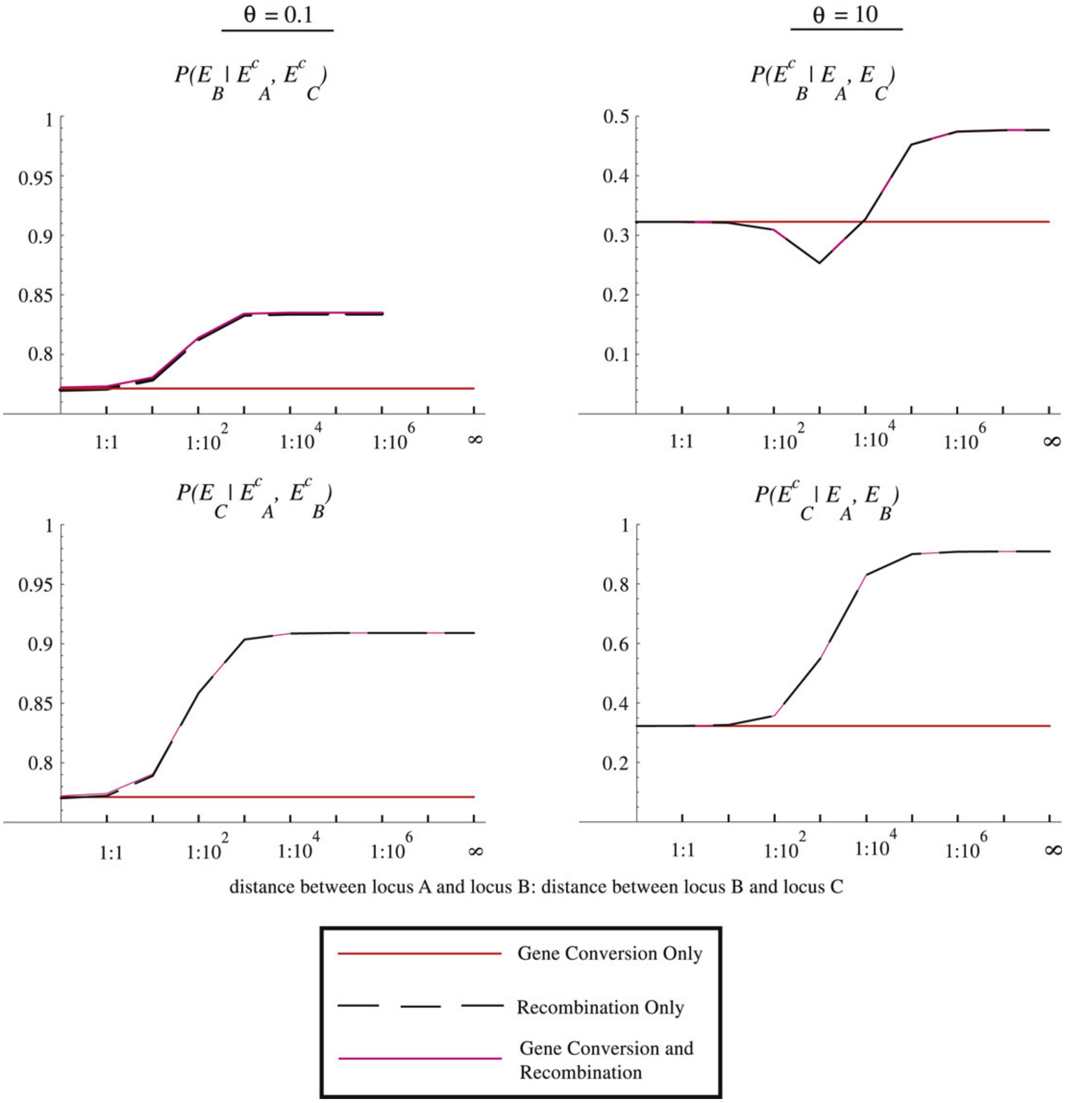


Fig. 8. When the distance between locus A and B is equal and the distance between locus B and C increases in tenfold increments. Four probabilities are examined under three conditions: the gene conversion only scenario (thin line), the recombination only scenario (broken, black line), and the recombination and gene conversion scenario (thin line between thick, broken line). $\kappa = 0.01$.

the denominator also increases which, in turn, decreases the probability. In contrast, $P(E_B^c|E_A, E_C)$ is conditioned on no mutations at either locus A or locus C so as θ increases, the denominator decreases which should increase the probability.

Fig. 7 shows that the two probabilities are higher when gene conversion is available compared to when recombination alone is present. The largest difference between the probabilities produced by the two crossing over types, approximately equal to 0.13, occurs when $\theta = 10$ and $\rho = 20$ for $P(E_B^c|E_A, E_C)$. In comparison, for $P(E_B|E_A^c, E_C^c)$ the largest probability difference found within the range of parameters estimated for human data in our model is 0.024 when $\theta = 0.1$ and $\rho = 2.0$.

As the rate of crossing over increases, it becomes easier to separate the histories at the middle and outer loci even when only recombination (and no gene conversion) is present. As a consequence, for any given value of θ, there is an increase in the two probabilities (Eqs. (3) and (4)) as the rate of crossing over increases. When the rate of crossing over is very small or very large, of both recombination and gene conversion produce the same probability within Eqs. (3) and (4); thus, the lines in each graph of Fig. 7 meet at two points. The probabilities calculated with recombination (and no gene conversion) align with those that were found using the ms program (Hudson, 2002) with recombination and the same mutation rates.

Fig. 8 shows that the four examined probabilities (Eqs. (3)–(6)) produce constant values with gene conversion (and no recombination). The same probabilities calculated with recombination, in contrast, increase as the distance between each locus increases. Since some of the largest differences seen in probabilities in Fig. 7 occurred for $P(E_B|E_A^C, E_C^C)$ when $\theta = 0.1$ and for $P(E_B^C|E_A, E_C)$ when $\theta = 10$, those were the mutation rates adopted for the calculation of the respective probabilities under the condition of increasing distance between locus B and locus C. The mutation rate, when calculating the four probabilities, was kept low when the probability was conditioned on two mutations (Eqs. (4) and (6)) and it was forced to be high when the probability was conditioned on one mutation (Eqs. (3) and (5)).

In Fig. 8 as the distance between locus B and locus C increases the four probabilities, with recombination, increase. By definition, as distance between two loci increases there are more recombination events which occur between the two loci. The difficulty of obtaining the series of events necessary to separate the history of the middle locus from the outer loci is lessened and so $P(E_B|E_A^c, E_C^c)$ and $P(E_B^c|E_A, E_C)$ increase. The difference in probabilities is particularly noticeable when there is 1000–100 000 times more distance between locus B and locus C then between locus A and locus B, at which point the probabilities have achieved their equilibrium values.

The far right of Fig. 8 demonstrates that when there is an infinite amount of crossing over, the model collapses to a two-locus case. Thus, $P(E_B^c|E_A, E_C) = P(E_B^c|E_A)$. When there is an infinite amount of recombination between locus B and C, locus C has no relevance on the probability of locus B being IBD or not and vice versa. For Eqs. (5) and (6), the equations, which describe the state of locus C, become independent of the other two loci and reduce to a one-locus equation. For instance, when $\theta = 10$, Eq. (5) reduces to $\theta/(1+\theta) = 1/11$ and Eq. (6) reduces to $\theta/(1+\theta) = 10/11$.

4. Discussion

We used a coalescent approach to build a matrix that chronicles the histories of three neighboring loci. The matrix has seven absorbing states that describe the possible patterns of coalescence and crossing over among three loci in a sample of size two. At equilibrium, the probabilities of these absorbing states provide the probability of IBD for three loci under a range of parameter values.

One of the absorbing states (state 34 above) gives the probability that the two outer loci (A and C above) have a shared coalescence history that differs from the middle locus, namely that locus B coalesces prior to the other two. Gene conversion is more likely than recombination to produce state 34 (Figs. 5 and 6) because it can separate the ancestry of locus B while preserving the shared history of locus A and locus B. Contrary to intuition, over a wide range of parameter values, we have shown that conditional patterns of IBD meant to capture this difference may not be very useful in distinguishing gene conversion from recombination (Figs. 7 and 8).

Although there is some promise in Figs. 4–6, the patterns of IDB at three loci in a sample of size two are not strikingly different under gene conversion than under recombination. This can be attributed to the fact that 2/3 of the time, when it acts upon either locus A or locus C, gene conversion looks identical to recombination. Our ability to distinguish gene conversion from recombination is further hampered by the fact that we cannot directly observe patterns of genetic ancestry. For this we depend on mutation. The somewhat discouraging picture painted by Figs. 7 and 8 reflects the fact that for realistic (small) values of θ, the strong tendency towards IBD can obscure underlying patterns of ancestry.

Our analysis is, however, limited by our consideration of only three loci from two sampled sequences. Statistics based on larger numbers of loci and/or larger samples could be more powerful. Using simulations, Wiuf and Hein (2000) note that the power to detect gene conversion from sequence data is low for small samples, but increases with sample size. Wall (2004) further suggests that accurate estimation of recombination and gene conversion requires data from at least ten independent loci. Our work supports these statements, but also indicates that analytical treatment of more than three loci from samples larger than two will be difficult.

Acknowledgments

The authors thank Christina Muirhead for her very useful advice on Mathematica, Bjarki Eldon for his patiently dispensed pointers on LaTeX and Alexander Platt and Daven Presgraves for many fruitful discussions.

Appendix

The probability of IBD for three and two loci were determined by adding up the seven (or three, for the two-locus case) absorbing states in the transition matrix at equilibrium under various parameter values for the first row of the matrix. The first row gives transition starting from two chromosomes, each with three loci, and tracking them backwards in time until coalescence or mutation.

A.1. The derivation of the two-locus and three-locus probabilities involved in the determination of $P(E_B|E_A^c, E_C^c)$:

$$P(E_B|E_A^c, E_C^c) = \frac{P(E_A^c, E_B, E_C^c)}{P(E_A^c, E_C^c)}. \tag{A.1}$$

The two-locus probability is the following:

$$P(E_A^c, E_C^c) = P(E_C^c) - P(E_A, E_C^c) \tag{A.2}$$

$$P(E_C^c) = 1 - P(E_C) \tag{A.3}$$

$$P(E_A, E_C^c) = P(E_A) - P(E_A, E_C). \tag{A.4}$$

Substitute (A.3) and (A.4) into (A.2). This gives the two-locus probability of being not-identical-by-descent as:

$$P(E_A^c, E_C^c) = 1 - P(E_C) - P(E_A) + P(E_A, E_C). \tag{A.5}$$

The three-locus probability is derived according to the following:

$$P(E_B, E_A^c, E_C^c) = P(E_B, E_C^c) - P(E_B, E_A, E_C^c) \tag{A.6}$$

$$P(E_B, E_C^c) = P(E_B) - P(E_B, E_C) \tag{A.7}$$

$$P(E_B, E_A, E_C^c) = P(E_A, E_B) - P(E_A, E_B, E_C). \tag{A.8}$$

Substitute (A.7) and (A.8) into (A.6). This gives:

$$\begin{aligned} P(E_B, E_A^c, E_C^c) = {} & P(E_B) - P(E_B, E_C) - P(E_A, E_B) \\ & + P(E_A, E_B, E_C). \end{aligned} \tag{A.9}$$

Substitute (A.5) and (A.9) into (A.1). This gives:

$$P(E_B|E_A^c, E_C^c) = \frac{P(E_B) - P(E_B, E_C) - P(E_A, E_B) + P(E_A, E_B, E_C)}{1 - P(E_C) - P(E_A) + P(E_A, E_C)}. \quad \text{(A.10)}$$

A.2. The derivation of the two-locus and three-locus probabilities involved in the determination of $P(B^c|A, C)$:

$$P(E_B^c|E_A, E_C) = \frac{P(E_A, E_B^c, E_C)}{P(E_A, E_C)}. \quad \text{(A.11)}$$

The two-locus probability is the following:

$$P(E_A, E_C). \quad \text{(A.12)}$$

The three-locus probability is derived according to the following:

$$P(E_A, E_B^c, E_C) = P(E_A, E_C) - P(E_A, E_B, E_C) \quad \text{(A.13)}$$

Substitute (A.12) and (A.13) into (A.11). This results in:

$$P(E_B^c|E_A, E_C) = \frac{P(E_A, E_C) - P(E_A, E_B, E_C)}{P(E_A, E_C)}.$$

References

Hein, J., Schierup, M., Wiuf, C., 2005. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory, 1st edition. Oxford University Press, USA.

Nordborg, M., 2001. Coalescent theory. In: The Handbook of Statistical Genetics. Wiley, Chichester, UK (Chapter 7).

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

Ardlie, K., Liu-Cordero, S.N., Eberle, M.A., Daly, M., Barett, J., Winchester, E., Lander, E.S., Kruglyak, L., 2001. Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. American Journal of Human Genetics 69, 582–589.

Frisse, L., Hudson, R.R., Bartoszewicz, A., Wall, J.D., Donfack, J., DiRienzo, A., 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium. American Journal of Human Genetics 69, 831–843.

Hernández-Sánchez, J., Haley, C.S., Woolliams, J.A., 2006. Prediction of IBD based on population history for fine gene mapping. Genetics Selection Evolution 38, 231–252.

Hill, W.G., Hernández-Sánchez, J., 2007. Prediction of multi-locus Identity-by-Descent. Genetics 176, 2307–2315.

Hill, W.G., Weir, B.S., 2007. Prediction of multi-locus inbreeding coefficients and relation to linkage disequilibrium in random mating populations. Theoretical Population Biology 172, 179–185.

Hudson, R.R., 2001. Two locus sampling distributions and their application. Genetics 159, 1805–1817.

Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model. Bioinformatics 18, 337–338.

Innan, H.B., Padhukasahasram, B., Nordborg, M., 2003. The pattern of polymorphism on human chromosome 21. Genome Research 13, 1158–1168.

Jeffreys, A., May, C., 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nature Genetics 36, 151–156.

Jeffreys, A.J., Kauppi, L., Neumann, R., 2001. Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics 29, 217–222.

Jorde, L.B., 2005. Where we're hot, they're not. Science 308, 60–62.

Malecot, G., 1975. Heterozygosity and relationship in regularly subdivided populations. Theoretical Population Biology 8, 212–241.

Padhukasahasram, B., Marjoram, P., Nordborg, M., 2004. Estimating the rate of gene conversion on human chromosome 21. American Journal of Human Genetics 75, 386–397.

Pritchard, J., Przeworski, M., 2001. Linkage disequilibrium in humans: Models and data. American Journal of Human Genetics 69, 1–14.

Przeworski, M., Wall, J.D., 2001. Why is there so little intragenic linkage disequilibrium in humans? Genetical Research 77, 143–151.

Ptak, S.E., Voelpel, K., Przeworski, M., 2004. Insights into recombination from patterns of linkage disequilibrium in humans. Genetics 167, 387–397.

Song, Y.S., Ding, Z., Gusfield, D., Langley, C.H., Wu, Y., 2006. Algorithms to distinguish the role of gene-conversion from single-crossover recombination in the derivation of SNP sequences in populations. Lecture Notes in Computer Science 3909, 231–245.

Stahl, F.W., 1994. The Holliday junction on its thirtieth anniversary. Genetics 138, 241–246.

Strobeck, C., Morgan, K., 1978. The effect of intragenic recombination on the number of alleles in a finite population. Genetics 88, 829–844.

Wall, J.D., 2004. Estimating recombination rates using three-site likelihood. Genetics 167, 1461–1473.

Wiuf, C., Hein, J., 2000. The coalescent with gene conversion. Genetics 155, 451–462.

Wolfram Research, Inc. 2002. Mathematica, Version 4.2. Champaign, IL.