J Hum Genet (2007) 52:738–746
DOI 10.1007/s10038-007-0176-8

ORIGINAL ARTICLE

Identifying haplotype block structure using an ancestor-derived model

Hironori Fujisawa · Minoru Isomura · Shinto Eguchi · Masaru Ushijima · Satoshi Miyata · Yoshio Miki · Masaaki Matsuura

Received: 10 April 2007 / Accepted: 22 June 2007 / Published online: 18 July 2007

Abstract Recently, haplotype-based association studies have become popular for detecting disease-related or drug-response-associated genes. In these studies, it has been gradually recognized that a haplotype block structure is important. A rational and automatic method for identifying the haplotype block structure from SNP data has been desired. We have developed a new method using an ancestor-derived model and the minimum description length principle. The proposed method was applied to real data on the *TAP2* gene in which a recombination hotspot was previously reported in human sperm data. The proposed method could identify an appropriate haplotype block structure, while existing methods failed. The performance of the proposed method was also investigated in a simulation study. The proposed method presented a better performance in real data analysis and the simulation study than existing methods. The proposed method was powerful from the viewpoint of hotspot sensitivity and was robust to mutation except at the edge of a sequence.

Keywords Ancestral haplotype · Haplotype block · Linkage disequilibrium · Minimum description length principle · Recombination hotspot

Electronic supplementary material The online version of this article (doi:10.1007/s10038-007-0176-8) contains supplementary material, which is available to authorized users.

H. Fujisawa (✉) · S. Eguchi
The Institute of Statistical Mathematics,
Tokyo 106-8569, Japan
e-mail: fujisawa@ism.ac.jp

M. Isomura · M. Ushijima · S. Miyata · Y. Miki · M. Matsuura
Genome Center, Japanese Foundation for Cancer Research,
Tokyo 135-8550, Japan

Introduction

Genetic markers based on variation in the human genome sequence play powerful roles in identifying disease-related or drug-response-associated genes. Single-nucleotide polymorphisms (SNPs) are considered to be promising genetic variations because of their abundant distribution in the human genome. Many association studies have been done using genotypes of SNPs and haplotypes restored by them. In these studies, it has been gradually recognized that identifying a haplotype block structure is an important way to narrow down candidate loci.

On chromosome 5q31, Daly et al. (2001) reported a block structure in which the SNPs exhibit strong linkage disequilibrium (LD) within each block, whereas SNPs exhibit a sudden decay of the LD between blocks. Such a structure is now referred to as a haplotype block structure and the position of sudden decay is referred to as a recombination hotspot. They furthermore pointed out that most haplotypes could be regarded as recombinants of only a few common haplotypes, which implies a low haplotype diversity. Hereafter, such common haplotypes are referred to as ancestral haplotypes.

One important research topic is to develop a rational and automatic method for identifying the underlying haplotype block structure. The methods proposed in the past can be classified into four types. The first type aims at achieving a low haplotype diversity. Daly et al. (2001) targeted locally minimal haplotype diversity based on haplotypic heterozygosity. Patil et al. (2001) focused on haplotype-tagging SNPs. This standpoint was extended by Zhang et al. (2002a, 2002b, 2003). The second type aims at detecting a recombination hotspot. Gabriel et al. (2002) and Wang et al. (2002) used the coefficient of linkage disequilibrium (LD) and the four-gamete test statistic (Hudson 1985) as

indices of recombination, respectively. The first two types were reexamined by Ding et al. (2005). The third type is an attempt to combine the above two ideas (Zhu et al. 2003; Kamatani et al. 2004). The fourth type is based on a statistical model and the minimum description length (MDL) principle (Rissanen 1978), which manages to handle both haplotype diversity and recombination hotspots. Some methods were proposed by Koivisto et al. (2003), Greenspan and Geiger (2003), and Jojic et al. (2004). However, their models do not use a special probability structure related to recombination events, which is well-known in statistical genetics.

Anderson and Novembre (2003) incorporated the special probability structure into a Markov model, which was one of the key points of their paper. They showed that their idea worked remarkably well in comparison with past ideas used in simulation studies and real data analyses. Their method is referred to as the "AN method" in this paper. Anderson and Slatkin (2004) analyzed the 5q31 data in more detail using the AN method. However, as described later, the AN method appears to be overly sensitive to mutation.

To overcome this issue, this paper proposes a new method for identifying the haplotype block structure. This paper constructs a statistical model as follows. The most important mechanism for generating the haplotype block structure is the existence of ancestral haplotypes and recombination hotspots. We first incorporate a probability structure for the recombination event with ancestral haplotypes into the central part of the statistical model, which is referred to as an ancestor-derived model, but the resulting model can only handle recombinants of ancestral haplotypes. The remaining problem is to model the remaining nonrecombinant haplotypes. Note that they are not interpreted by a simple genetic property (e.g., mutation) and their frequencies are small in general, so we adopt a simple full frequency model with no structure to avoid extra complexity of the model. The statistical model proposed is a mixture of the ancestor-derived model and the simple full frequency model. The optimal model is determined by the MDL principle. The proposed method was implemented in the ADBlock software.

The proposed method was applied to two real datasets. One is a locus in the *TAP2* gene on chromosome 6p21.3, in which a recombination hotspot was biologically identified by analyzing sperm DNA (Jeffreys et al. 2000). The genotype data were provided by the Japanese Foundation for Cancer Research (JFCR). The proposed method could identify an appropriate haplotype block structure, but some existing methods (based on the AN method, Gabriel et al. (2002), and PHASE) failed. The other is the 5q31 data analyzed by Daly et al. (2001). The haplotype block structure identified by the proposed method was similar to the conventional structure. The performance of the proposed method was also investigated in a simulation study.

Materials and methods

Ancestor-derived model

First, we define some notation. Let the numbers of SNPs and observed haplotypes be denoted by L and N, respectively. The ith observed haplotype can be expressed as $h_i = (h_{i1}, ..., h_{iL})$ for $i = 1, ..., N$, where $h_{il} \in \{0,1\}$ because the SNP is biallelic. The suffix i is sometimes omitted for simplicity. The block structure $\mathcal{B} = \{B^{(1)}, \ldots, B^{(K)}\}$ can be expressed as the partition of SNPs, where K is the number of blocks, $B = \{1, ..., L\} = B^{(1)} \cup ... \cup B^{(K)}$, and $B^{(k)} = \{l_{k-1} + 1, ..., l_k\}$ is the set of adjacent SNPs with $l_0 = 0$ and $l_K = L$. The haplotype corresponding to the above partition is denoted by $h = (h^{(1)}, ..., h^{(K)})$, where $h^{(k)} = (h_{l_{k-1}+1}, \ldots, h_{l_k})$ is the partial haplotype on block $B^{(k)}$. Let $\mathcal{G} = \{g_1, \ldots, g_A\}$ be the set of ancestral haplotypes and q_a the frequency probability of g_a, where $\sum_{a=1}^{A} q_a = 1$. Assume that the observed haplotype h is a recombinant haplotype, more precisely, $h = \left(g_{a_1}^{(1)}, \ldots, g_{a_K}^{(K)}\right)$ for some $(a_1, ..., a_K)$.

Let $R = (R_1, ..., R_{K-1})$ be the latent indicator of whether the recombination event happens or not, where $R_k = 1$ if the recombination event happens between two blocks $B^{(k)}$ and $B^{(k+1)}$ and $R_k = 0$ otherwise. Let the recombination rate be denoted by $\lambda_k = Pr\ (R_k = 1)$. Assume that the R_k's are independent variables. The probability of R is expressed as

$$Pr(R = r) = \prod_{k=1}^{K-1} \lambda_k^{r_k} (1 - \lambda_k)^{1-r_k}.$$

Suppose that R is given in the following. Another partition of SNPs is denoted by $B = B^{[1]} \cup \cdots \cup B^{[K^*]}$, where $K^* = \sum_{k=1}^{K-1} R_k + 1$, $B^{[k^*]}$ is the union of some adjacent $B^{(k)}$'s, the recombination events happen only between $B^{[k^*]}$'s and not within each $B^{[k^*]}$. A relationship between two block partterns is displayed in Fig. 1. The haplotype corresponding to the above partition is denoted by $h = \left(h^{[1]}, \ldots, h^{[K^*]}\right)$. The partial haplotype $h^{[k^*]}$'s derive from g_a's because the haplotype is a recombinant of ancestral haplotypes. This correspondence can be expressed as the indicator variable $C = (C^{[1]}, \ldots, C^{[K^*]})$, where $C^{[k^*]} = a$ indicates that $h^{[k^*]}$ derives from g_a. Note that the variable C depends on the structure of R. The conditional probability of C given R is defined as

$$Pr(C = c|R = r) = \prod_{k^*=1}^{K^*} \prod_{a=1}^{A} q_a^{I(c^{[k^*]}=a)},$$

SNP number	l	$1 \ldots l_1$	$l_1+1 \cdots l_2$	$l_2+1 \cdots l_3$	$l_3+1 \cdots l_4$	$l_4+1 \cdots l_5$
Haplotype	h_l	$h_1 \ldots h_{l_1}$	$h_{l_1+1} \cdots h_{l_2}$	$h_{l_2+1} \cdots h_{l_3}$	$h_{l_3+1} \cdots h_{l_4}$	$h_{l_4+1} \cdots h_{l_5}$
Original partition						
Block structure	$B^{(k)}$	$B^{(1)}$	$B^{(2)}$	$B^{(3)}$	$B^{(4)}$	$B^{(5)}$
Haplotype	$h^{(k)}$	$h^{(1)}$	$h^{(2)}$	$h^{(3)}$	$h^{(4)}$	$h^{(5)}$
Recombination	R_k	$R_1=0$	$R_2=1$	$R_3=0$	$R_4=0$	
Suffixes change after R is given.						
Block structure	$B^{[k^*]}$	$B^{[1]}$		$B^{[2]}$		
Haplotype	$h^{[k^*]}$	$h^{[1]}$		$h^{[2]}$		

Fig. 1 Illustrative example of notation in the case where $K = 5$ and $K^* = 2$. SNP number l ranges from $1 = l_0 + 1$ to $L = l_5$. The component of haplotype is given by $h_l = 0,1$. The kth block and partial haplotype are denoted by $B^{(k)} = \{l_{k-1} + 1, \ldots, l_k\}$ and $h^{(k)} = (h_{l\_k-1} + 1, \ldots, h_{l\_k})$, respectively. The recombination event is expressed as $R_k = 1$ if it happens between $B^{(k)}$ and $B^{(k+1)}$ and $R_k = 0$ otherwise. After R is given as above, the new blocks and partial haplotypes are given as follows: $B^{[1]} = B^{(1)} \cup B^{(2)}$. $B^{[2]} = B^{(3)} \cup B^{(4)} \cup B^{(5)}$. $h^{[1]} = (h^{(1)},h^{(2)})$. $h^{[2]} = (h^{(3)},h^{(4)}, h^{(5)})$

where $I(\mathcal{A})$ is one if $\mathcal{A}$ is true and zero otherwise. Consequently, the probability that the latent variable is observed is given by $Pr\,(R = r, C = c) = Pr\,(R = r)\,Pr\,(C = c \mid R = r)$, which is called a complete model (McLachlan and Peel 2000).

Recall the original problem. The event where the haplotype h is observed is expressed as

$$\mathcal{F}_h = \left\{(R, C)|h = \left(g^{[1]}_{C^{[1]}}, \ldots, g^{[K^*]}_{C^{[K^*]}}\right)\right\},$$

which implies a frequency probability

$$Pr(H = h) = \sum_{(r,c)\in\mathcal{F}_h} Pr(R = r, C = c),$$

which is referred to as the ancestor-derived model. This model is determined by the haplotype block structure, $\mathcal{B}$, and the set of ancestral haplotypes, $\mathcal{G}$. The parameters of the model consist of the recombination rate, $\lambda = (\lambda_1,\ldots, \lambda_{K-1})$, and the frequency of ancestral haplotype, $q = (q_1,\ldots,q_A)$.

To fully understand the ancestor-derived model, we will illustrate the simple case where the number of SNPs is five, $L = 5$, the block structure is given by $B^{(1)} = \{1,2,3\}$ and $B^{(2)} = \{4,5\}$, and the ancestral haplotypes are given by $g_1 = (0,0,0,0,0)$ and $g_2 = (1,1,1,1,1)$. Suppose that the observed haplotype is $h = (0,0,0,1,1)$. The recombination event certainly happens between two blocks, $R = 1$. The observed haplotype is a result of the connection between two partial descendents, more precisely, $h = (h^{[1]}, h^{[2]}) = (g_1^{[1]}, g_2^{[2]})$ and $C = (1, 2)$. It therefore follows that $\mathcal{F}_h = \{(1, (1, 2))\}$ and

$$\begin{aligned} Pr(H = h) &= Pr(R = 1, C = (1, 2)) \\ &= Pr(R = 1)Pr(C = (1, 2)|R = 1) = \lambda q_1 q_2. \end{aligned}$$

Suppose that the observed haplotype is $h = (0,0,0,0,0)$. If no recombination event happens, $R = 0$, then only one block is present, $K^* = 1$, and the observed haplotype is a direct descendent of the ancestral haplotype g_1, $C = 1$. Consider the case where the recombination event happens, $R = 1$. The number of blocks is two, $K^* = 2$. The observed haplotype is a result of the connection between two partial descendents, more precisely, $h = (h^{[1]}, h^{[2]}) = (g_1^{[1]}, g_1^{[2]})$ and $C = (1,1)$. It therefore follows that $\mathcal{F}_h = \{(0, 1), (1, (1, 1))\}$ and

$$\begin{aligned} Pr(H = h) &= Pr(R = 0, C = 1) + Pr(R = 1, C = (1, 1)) \\ &= (1 - \lambda)q_1 + \lambda q_1^2. \end{aligned}$$

The general case can be extended in a similar way.

Mixture model and parameter estimation

The ancestor-derived model has been constructed to handle recombinant haplotypes. The remainder of the problem is to model the remaining nonrecombinant haplotypes. As described in the "Introduction," a simple full frequency model is applied to nonrecombinant haplotypes. Let $\mathcal{U} = \{u_1, \ldots, u_D\}$ be the set of distinct nonrecombinant haplotypes and p_d be the frequency probability of u_d, where $\sum_{d=1}^{D} p_d = 1$. The full frequency model is given by $Pr(H = h) = \prod_{d=1}^{D} p_d^{I(h=u_d)}$. The whole model proposed is a mixture of the ancestor-derived model and the simple full frequency model, with a mixing proportion ω.

The parameter of the mixture model, $\theta = (\lambda, q, p, \omega)$, can be estimated by the maximum likelihood principle. Let the maximum likelihood estimate of θ be denoted by $\hat{\theta}$. Note that the observed haplotype belongs to either of two underlying models and never to both simultaneously. Let the set of recombinant haplotypes be denoted by $\{h_1, \ldots, h_n\}$. It is clear that $\hat{\omega} = n/N, \hat{p}$ is simply given by the observed frequency, and $\hat{\xi} = (\hat{\lambda}, \hat{q})$ is the maximizer of $\sum_{i=1}^{n} \log Pr(H = h_i; \xi)$ with respect to $\xi = (\lambda,q)$, which can be obtained through the EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997). The detailed derivation of the algorithm is given in the "Electronic supplementary material" (ESM).

When the frequency probability, $Pr\ (H = h_i)$, is evaluated, many calculations may be needed, because the size of $\mathcal{F}_h$ is an exponential order of the number of blocks; more precisely, $(A + 1)^{K-1}$. This task can be reduced up to a polynomial order of low degree by virtue of the DP algorithm. The detailed derivation of the algorithm is given in ESM. In particular, its effectiveness appears in the parameter estimation because the frequency probability is calculated for each iteration step of the EM algorithm.

Code length of the model

The mixture model is determined by the haplotype block structure, $\mathcal{B}$, the set of ancestral haplotypes, $\mathcal{G}$, the set of the distinct nonrecombinant haplotypes, $\mathcal{U}$, and the resulting probability structure. By summing their code lengths, we can obtain the whole code length of the model. In the following, the code length is calculated fully. The optimal model is selected as the minimizer of the code length by the MDL principle (Rissanen 1978; Hansen and Yu 2001).

The haplotype block structure, $\mathcal{B}$, can be determined by the number of blocks, K, and the corresponding block patterns. We see that K ranges from 1 to L and that the number of possible block patterns is $v_{L,K} = (L - 1)!/(K - 1)!(L - K)!$. Hence, the necessary code length of $\mathcal{B}$ is given by $\log L + \log v_{L,K}$.

Next, consider two sets of haplotypes, $\mathcal{G}$ and $\mathcal{U}$, which are determined by their sizes and haplotypes. The sizes of the two sets are expressed by A and D, respectively. The size of the combined set, $A + D$, ranges from 1 to 2^L, whose necessary code length is $\log 2^L = L$. In this situation, it should be noted that the sum $A + D$ is known, although A and D are unknown. We see that A ranges from 1 to $A + D$, whose necessary code length is $\log(A + D)$. Therefore, the sizes of $\mathcal{G}$ and $\mathcal{U}$ can be coded by the length $L + \log(A + D)$. Note that a haplotype can be coded by the length L because each component is biallelic and the number of components is L. The necessary code length of the haplotypes of $\mathcal{G}$ and $\mathcal{U}$ is given by $(A + D)L$, because the number of haplotypes is $A + D$. Consequently, the necessary code length of $\mathcal{G}$ and $\mathcal{U}$ is given by $(A + D + 1)L + \log(A + D)$.

The code length of the probability structure is approximated by the negative maximum log-likelihood plus $\log N/2$ times the number of parameters. The total number of parameters is $(A + D - 2) + (K - 1) + 1 = A + D + K - 2$. It therefore follows that the whole code length of the model is given by

$$\psi = \log L + \log v_{L,K} + (A + D + 1)L + \log(A + D) - \sum_{i=1}^{N} \log Pr(H = h_i; \hat{\theta}) + (A + D + K - 2)\log N/2.$$

There are two problems when the MDL principle is applied directly to the real data. The existence of a very rare haplotype may lead to an erroneous conclusion, which is known as oversensitivity to outliers in a statistical field. In general, the number of distinct observed haplotypes is large; hence, the number of candidates for $\mathcal{G}$ is extraordinarily large. The former problem can be avoided by avoiding the use of a very rare haplotype. The latter problem can be overcome by using prior information that the ancestral haplotype has a much larger frequency. For a detailed procedure, see the "Discussion" section and the "ESM".

Comparison of the ancestor-derived model with the Markov model

The ancestor-derived model can interpret recombinant haplotypes using a smaller number of parameters than the Markov model constructed by Anderson and Novembre (2003). The Markov model has a flexible transition structure with A strength parameters to give a stronger connection between A ancestral haplotypes at each recombination hotspot. The ancestor-derived model has only one parameter used as a recombination rate at each recombination hotspot. It follows that the extra number is $(A - 1)(K - 1)$, which implies that the proposed method has a greater capability of identifying the correct haplotype block structure. This was verified in a simulation study.

More than one generation change

Let $Pr^{[m]}(H = h)$ be the probability that the haplotype h is observed at the mth generation as a recombinant of the ancestor haplotypes of the 0th generation. The case $m = 1$ corresponds to the ancestor-derived model. For simplicity, consider the case $K = 2$. Let the recombination rate at each generation change be denoted by ζ. It can be shown that

$$Pr^{[m]}(H = h) = (1 - \zeta)^m Pr^{[0]}(H = h) + \{1 - (1 - \zeta)^m\} Pr^{[0]}\left(H^{(1)} = h^{(1)}\right) Pr^{[0]}\left(H^{(2)} = h^{(2)}\right).$$

The derivation is given in the "ESM." It can be easily verified that this is completely the same as the ancestor-derived model where the recombination rate λ is replaced by $\{1 - (1 - \zeta)^m\}$. Thus, the ancestor-derived model has a reasonable understanding, even for more than one generation. Furthermore, since ζ is sufficiently small in general, it holds that $\lambda \approx m\zeta$, so that the recombination rate λ of the ancestor-derived model approximately corresponds to the true recombination rate times the number of generation changes.

Table 1 Locations of the SNPs in the *TAP2* gene

	dbSNP ID	Position
SNP 1	rs2856993	32899380
SNP 2	rs1044043	32901959
SNP 3	rs241453	32904204
SNP 4	rs241429	32911817
SNP 5	rs2239701	32913026
SNP 6	rs2071465	32913447
SNP 7	rs2071544	32914098
SNP 8	rs2071552	32914438
SNP 9	rs3763366	32915423
SNP 10	rs3763365	32915430
SNP 11	rs3763364	32915497

Table 2 Two underlying sets of ancestral haplotypes and their frequencies (in parentheses) for simulation

$A = 2$	$A = 4$
0000000000 (0.8)	0110101010 (0.4)
1111111111 (0.2)	0001001000 (0.3)
	1001011101 (0.2)
	0111100000 (0.1)

A is the number of ancestral haplotypes

Genotyping of SNPs in the *TAP2* gene

As a part of the pharmacogenomic research being performed at JFCR, the genotyping of SNPs in the *TAP2* gene was conducted. After obtaining written informed consents, 15 ml of peripheral blood were collected from 719 Japanese individuals and DNA were isolated. Genotypes of 11 SNPs in the *TAP2* gene were determined by Invader assay. The 11 SNPs used in this study are listed in Table 1.

Results

Simulation study

An artificial sample was generated as follows. Two underlying sets of ancestral haplotypes and their frequencies are given in Table 2. The ancestral haplotypes in the case $A = 4$ are the same as blocks 4 and 5 of Daly et al. (2001). The recombination hotspot was set between SNPs 5 and 6 with the recombination rate $\lambda = 0.1, 0.3$. We drew upon Daly et al. (2001) to set various parameter values. The background recombination rate was set to be 10^{-3}. Each setting uniquely determines the ancestor-derived model. We randomly sampled n haplotypes from the model for $n = 200$, 500, 1,000 and furthermore $n = 2{,}000$ in the case $A = 4$, and then we exposed each haplotype component to mutation with rate $\mu = 0.001, 0.01, 0.03$. Note that their mutation rates correspond to the cases where the rates of unexposed haplotype are $(1 - \mu)^{10} \approx 1 - 10\mu = 0.99$, 0.9, 0.7, respectively.

Based on 100 samples, the proposed method and the AN method were compared in regard to the hotspot and nonhotspot sensitivities, which are the fractions of the hotspots and nonhotspots that are judged correctly. The case where the estimated recombination rate was less than 0.03 was neglected in the proposed method, because such a case does not present sufficient evidence of a recombination hotspot. The AN method was carried out using the MDBlocks software provided by Anderson and Novembre (2003).

The results for the cases where $\lambda = 0.1$ and $\mu = 0.01$, 0.03 are displayed in Tables 3 and 4, where spot k indicates the position between SNPs k and $k + 1$. The hotspot and nonhotspot sensitivities correspond to spot 5 and the other spots, respectively. The case $\mu = 0.001$ presented almost 100% efficiency from the viewpoint of nonhotspot sensitivity (not shown). The results in the case $\lambda = 0.3$ were similar to those in the case $\lambda = 0.1$, but the hotspot sensitivities in the case $\lambda = 0.3$ (not shown) were larger than those in the case $\lambda = 0.1$ because the recombination rate was larger.

The proposed method was superior to the AN method from the viewpoint of hotspot sensitivity, for the reason described previously. For the cases where $(\mu, A, n) = (0.01, 2, 1000)$, $(0.03, 2, 500)$, $(0.03, 2, 1000)$, $(0.03, 4, 1000)$, and $(0.03, 4, 2000)$, the proposed method was more robust to mutation than the AN method. In particular, in the case $(\mu, A, n) = (0.03, 2, 1000)$, the AN method judged all of the spots as being hotspots, which was completely wrong. The AN method was sensitive to mutation when the number of mutation events was not small. We expect that the method is powerful when the sample size is large. However, the AN method is not always powerful when the sample size is large. On the other hand, the proposed method was less stable than the AN method from the viewpoint of nonhotspot sensitivity near the edge of a sequence. In the simulation study, it was shown that the proposed method is powerful from the viewpoint of hotspot sensitivity and is robust to mutation except at the edge of a sequence.

The proposed method often regarded a nonhotspot as a hotspot near the edge of a sequence. The reason for this is as follows. Suppose that the haplotype data includes two ancestral haplotypes 00... 0 and 11... 1 and mutant haplotypes (nonrecombinants) 10... 0 and/or 01... 1. Assume that the first spot is a hotspot. The maximum log-likelihood will decrease because of erroneous modeling, but the magnitude of the decrease may not be as large because the SNP is at

Table 3 Hotspot and nonhotspot sensitivities in the case where $\lambda = 0.1$ and $\mu = 0.01$

Spot	1	2	3	4	5	6	7	8	9
$A = 2$, $n = 200$									
ADBlock	75	100	100	100	45	100	100	100	86
MDBlocks	100	100	100	100	20	100	100	100	100
$A = 2$, $n = 500$									
ADBlock	55	100	100	100	97	100	100	100	75
MDBlocks	95	100	100	100	97	100	100	100	95
$A = 2$, $n = 1{,}000$									
ADBlock	78	100	100	100	97	100	100	100	85
MDBlocks	58	59	57	57	90	47	50	48	47
$A = 4$, $n = 200$									
ADBlock	96	100	100	97	39	100	100	97	100
MDBlocks	100	100	100	100	0	100	100	100	100
$A = 4$, $n = 500$									
ADBlock	72	94	100	100	99	100	100	76	100
MDBlocks	100	100	100	99	81	96	100	100	100
$A = 4$, $n = 1{,}000$									
ADBlock	36	85	100	92	99	99	99	46	100
MDBlocks	100	99	100	100	99	99	99	85	100
$A = 4$, $n = 2000$									
ADBlock	22	85	100	94	100	100	99	10	100
MDBlocks	100	100	100	100	100	100	11	3	99

Spot 5 is a hotspot and the other spots are nonhotspots. These are the fractions of hotspots and nonhotspots that are judged correctly, based on 100 random samples

Table 4 Hotspot and nonhotspot sensitivities in the case where $\lambda = 0.1$ and $\mu = 0.03$

Spot	1	2	3	4	5	6	7	8	9
$A = 2$, $n = 200$									
ADBlock	63	100	100	100	37	100	100	100	57
MDBlocks	98	100	100	100	3	100	100	100	100
$A = 2$, $n = 500$									
ADBlock	47	100	100	100	92	100	100	100	41
MDBlocks	44	44	42	44	66	47	49	49	49
$A = 2$, $n = 1{,}000$									
ADBlock	21	100	100	100	99	100	100	100	17
MDBlocks	0	0	0	0	100	0	0	0	0
$A = 4$, $n = 200$									
ADBlock	77	76	100	69	67	97	100	81	100
MDBlocks	100	100	100	100	2	100	100	100	100
$A = 4$, $n = 500$									
ADBlock	27	36	99	52	92	90	88	21	100
MDBlocks	100	96	100	100	49	92	98	85	100
$A = 4$, $n = 1{,}000$									
ADBlock	29	11	100	83	98	98	100	10	100
MDBlocks	83	70	68	100	72	72	36	8	92
$A = 4$, $n = 2000$									
ADBlock	41	12	100	94	100	100	99	10	100
MDBlocks	21	14	14	100	86	86	14	4	96

Spot 5 is a hotspot and the other spots are nonhotspots. These are the fractions of hotspots and nonhotspots that are judged correctly, based on 100 random samples

most biallelic. On the other hand, the necessary code length of $\mathcal{U}$ decreases because the mutant haplotypes can be expressed as recombinants of two ancestral haplotypes and then the mutant haplotypes are taken away from $\mathcal{U}$. Such a trade-off of code length could lead to a decrease in the whole code length, which would lead to an incorrect

haplotype block structure. A suspicious hotspot near the true hotspot was also explained in the same way.

The AN method yielded a great many suspicious hotspots when the number of mutation events was not small, as described already. The reason for this is very similar to that for the proposed method, and furthermore the decrease in the maximum log-likelihood will be smaller than that for the proposed method, because the Markov model is very flexible, so that it seems that the AN method is more sensitive to mutation at all spots than the proposed method.

Analysis of data for the *TAP2* gene

Jeffreys et al. (2000) biologically identified a recombination hotspot in the *TAP2* gene during their investigation of sperm DNA. We sampled the genotypes of 11 SNPs in the *TAP2* gene for 719 individuals. The SNPs used are listed in Table 1. The recombination hotspot is located between SNPs 3 and 4; in other words, at spot 3. We tested whether the proposed method could correctly identify the recombination hotspot and then compared the results from the proposed method with those from the AN method, the method proposed by Gabriel et al. (2002), and PHASE (http://www.stat.washington.edu/stephens/software.html).

Some genotypes remained undetermined due to incomplete reactions. The missing responses might be replaced with some alternatives, but uncertainty about the filled responses would lead to an unreliable conclusion. For this reason, the individual for whom more than 10% of the genotype was missing was not used. The haplotype data were restored from the genotype data by the Haplotyper (Niu et al. 2002), which is based on a model-based approach (Excoffier and Slatkin 1995) and Bayesian inference. The proposed method was applied to the restored haplotype data.

The top 5 code lengths and the corresponding haplotype block structures are given in Table 5. The smallest code length shows the most probable result. Compared with the biologically identified recombination hotspot, the proposed method could identify an appropriate haplotype block structure.

The second case corresponds to the one where no hotspot is present. The third and fourth cases imply two possibilities. One is the existence of a hotspot at spot 1, and the other is that the proposed method is not stable near the edge of the sequence. The latter reason would be correct, because the smallest code length treats spot 3 as the only hotspot. The last case would present suspicious adjacent hotspots, as described in the simulation study.

The identified ancestral haplotypes and their frequencies are given in Table 6. The first and seventh haplotypes are the same except for SNP 4, and the second to fourth and the eighth haplotypes are the same on the second block. They might be due to gene drift.

Table 5 Comparison of haplotype block structures (HBSs) identified by various methods in the *TAP2* gene

	HBS	Code length (ψ)
True	*0x0000000	
ADBlock	00x0000000	7054.015
	0000000000	7082.393
	x000000000	7084.261
	x0x0000000	7090.830
	00xx000000	7152.397
MDBlocks	x00xx00000	
Gabriel	xxxxxxx000	
PHASE	x0xxx00000	

x, 0, and * indicate a hotspot, a nonhotspot, and an unclear spot, respectively. It is unclear whether spot 1 in the true HBS is a hotspot or not

Table 6 Ancestral haplotypes and frequencies identified in the *TAP2* gene

Ancestral haplotype	Frequency
011-11010000	0.242
001-00101111	0.140
111-00101111	0.114
110-00101111	0.110
001-11101111	0.106
010-00111001	0.099
011-01010000	0.073
010-00101111	0.061
011-10110001	0.057
	1.000

Hyphen indicates the position of the hotspot identified

The AN method was also applied to the same restored haplotype data. The AN method detected two incorrect hotspots at spots 4 and 5, as shown in Table 5, which do not include the correct hotspot, although the sample size is sufficiently large. This erroneous result is similar to that observed in the simulation study.

The method proposed by Gabriel et al. (2002) was also applied to the data. A haplotype block between SNPs 8 and 11 was constructed, but many extra hotspots were also observed, as in Table 5. Such an erroneous result was also reported in Anderson and Novembre (2003).

PHASE was also applied to the data. PHASE gives estimates of the average background recombination rate and estimates of factors by which the recombination rate between two adjacent SNPs (i.e., at any spot) exceeds the background rate. These estimates are based on the general

model for varying recombination rate from Li and Stephen (2003). We obtained from the posterior distribution of the recombination parameters the 95% lower confidence limit of the factor by which the recombination rate exceeds the average background recombination rate. When it exceeds one, we judged the corresponding spot to be a recombination hotspot. We detected four recombination hotspots at spots 1, 3, 4, and 5, as in Table 5. Spots 4 and 5 were judged incorrectly.

Analysis of 5q31 data

Daly et al. (2001) reported a haplotype block structure on chromosome 5q31 with 103 SNPs. Using 258 individuals, Anderson and Novembre (2003) identified a haplotype block structure similar to that in Daly et al. (2001). The proposed method also identified a similar structure in a similar way to that described in the previous subsection. For details, see the "ESM."

Two sets of hotspots identified in Daly et al. (2001) and the proposed method were given by

$$\{8/9, 14/15, 24, 35, 40, 45, 76/77, 84/85, 91, 98\},$$
$$\{8, 15, 28, 37, 39, 44, 86, 90, 91, 92, 98\},$$

where, e.g., 8/9 means that either/both is/are the hotspot (because they did not clearly identify the hotspots in their paper). It should be noted that it is difficult to get a completely clear result because the sample size is not sufficiently large. The two sets are similar except for two different points. One is the missing hotspot 76/77 and the other is the existence of adjacent hotspots from spot 90 to 92. The missing hotspot 76/77 may be caused by the difference between the methods used for restoring haplotype data from genotype data. In our restored haplotype data, the number of recombination events illustrated in Daly et al. (2001) was only one. The reason for generating adjacent hotspots would be the same as that described in the simulation study.

Discussion

The ancestor-derived model is the same as a specific restricted Markov model and can handle recombinant haplotypes using a smaller number of parameters. This implies that the proposed method tends to be more powerful from the viewpoint of hotspot sensitivity than the AN method. This was verified in the simulation study. The robustness to mutation was also investigated in the simulation study. The proposed method was more robust to mutation than the AN method. The Markov model is very flexible, and its flexibility is often a favorable property for modeling. However, the combination of its flexibility and the MDL principle caused an overly sensitive response to mutation which erroneously yielded extra hotspots.

Some tuning parameters are prepared when the proposed method is applied (see "ESM" for details). One of them is the maximum of the observed number of haplotypes excluded in advance before the proposed method is applied, say M. As described already in "Code length of the model," we should avoid oversensitivity to outliers. Therefore, we set $M = 2$; that is, we excluded very rare haplotypes whose observed number is not more than $M = 2$ in advance. When such an approach was not used, we often identified suspicious hotspots. To illustrate that $M = 2$ is moderately appropriate, we carried out further simulations in a similar situation to that described the in the section "Simulation study" (results of these simulations are not shown here). As M was larger, the nonhotspot sensitivity tended to improve, but the hotspot sensitivity tended to get worse. The reason for the latter would be that some recombinant haplotypes were unnecessarily excluded from the analysis when they were accidentally rare. Remember that the hotspot and nonhotspot sensitivities were satisfactory except at the edge of a sequence when the proposed method was applied with $M = 2$. Therefore, we empirically recommend that when the proposed method is applied with $M = 2$, the results should be believed except for those for the edge of a sequence.

In the simulation study, the small recombination rate was neglected. In a real data analysis, it is recommended that the small recombination rate should be reviewed by comparing the resulting haplotype block structure with the original data.

The mutation event is an important factor to understand genome sequences, but is not incorporated into the statistical model in order to avoid extra model complexity. In fact, it is worth studying the incorporation of the mutation event into the model, but it would be very difficult to optimize this model because it causes a large number of calculations and oversensitivity to small frequencies relevant to the mutation event, which is well-known as outlier sensitivity in statistics.

The haplotype and the genotype have a latent relationship. We can prepare a new latent variable to introduce this latent relationship in addition to the latent variable (R,C). This new latent variable will enable us to construct an extended model to handle the genotype data directly. To make this idea feasible, it will be necessary to overcome another difficult calculation problem.

The proposed method identifies the optimal haplotype block structure but does not describe the statistical significance of the optimal structure compared to other candidates. A simple and feasible method for measuring the reliability of the optimal model may be to calculate how

often the optimal model is selected by the proposed method, based on bootstrap sampling (Efron and Tibshirani 1993). This is an important issue and further detailed research in this area is desirable.

Acknowledgments We would like to thank two reviewers for their helpful comments. This work was supported by a Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology, by grants from the New Energy and Industrial Technology Development Organization, and by ISM Project Research.

References

Anderson EC, Novembre J (2003) Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336–354

Anderson EC, Slatkin M (2004) Population-genetic basis of haplotype blocks in the 5q31 region. Am J Hum Genet 74:40–49

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38

Ding K, Zhou K, Zhang J, Knight J, Zhang X, Shen Y (2005) The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. Mol Biol Evol 22:148–159

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall/CRC, New York

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Greenspan G, Geiger D (2003) Model-based inference of haplotype block variation. In: Proc 7th Ann Int Conf on Research in Computational Molecular Biology—RECOMB, April 2003, Berlin, Germany, pp 10–13

Hansen M, Yu B (2001) Model selection and the principle of minimum description length. J Am Stat Assoc 96:746–774

Hudson RR, Kaplan N (1985) Statistical properties of the number of recombination events in the history of a sample of sequences. Genetics 111:147–164

Jeffreys AJ, Ritchie A, Neumann R (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum Mol Genet 22:725–733

Jojic N, Jojic V, Heckerman D (2004) Joint discovery of haplotype blocks and complex trait associations from SNP sequences. In: Proc 20th Conf on Uncertainty in Artificial Intelligence, July 2004, Banff, Canada, pp 286–292

Kamatani N, Sekine A, Kitamoto T, Iida A, Saito S, Kogame A, Inoue E, Kawamoto M, Harigai M, Nakamura Y (2004) Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP Maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. Am J Hum Genet 75:190–203

Koivisto M, Perola M, Varilo R, Hennah W, Ekelund J, Lukk M, Peltonen L, Ukkonen E, Mannila H (2003) An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. Pac Symp Biocomput 8:502–513

Li N, Stephens M (2003) Modelling linkage disequilibrium, and identifying recombination hotspots using snp data. Genetics 165:2213–2233

McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York

McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York

Niu TH, Qin ZHS, Xu XP, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Rissanen J (1978) Modeling by shortest data description. Automatica 14:465–471

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71:1227–1234

Zhang K, Calbrese P, Nordborg M, Sun F (2002) Haplotype block structure and its application to association studies: power and study designs. Am J Hum Genet 71:1386–1394

Zhang K, Deng M, Chen T, Waterman M, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339

Zhang K, Sun F, Waterman MS, Chen T (2003) Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. Am J Hum Genet 73:63–78

Zhu X, Yan D, Cooper RS, Luke A, Ikeda MA, Chang YP, Weder A, Chakravarti A (2003) Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. Genome Res 13:173–181