Genomics & Informatics Vol. 7(1) 1-12, March 2009

An Alternative Way of Constructing Ancestral Graphs Using Marker Allele Ages from Population Linkage Disequilibrium Information

Leeyoung Park*

Natural Science Research Institute, Yonsei University, Seoul 120-749, Korea

Abstract

An alternative way of constructing ancestral graphs, which is different from the coalescent-based approach, is proposed using population linkage disequilibrium (LD) data. The main difference from the existing method is the construction of the ancestral graphs based on variants instead of individual sequences. Therefore, the key of the proposed method is to use the order of allele ages in the graphs. Distinct from the previous age-estimation methods, allele ages are estimated from full haplotype information by examining the number of generations from the initial complete LD to the current decayed state for each two variants depending on the direction of LD decay between variants. Using a simple algorithmic procedure, an ancestral graph can be derived from the expected allele ages and current LD decay status. This method is different in many ways from previous methods, and, with further improvement, it might be a good replacement for the current approaches.

Keywords: allele age, ancestral graph, haplotypes, linkage disequilibrium, recombination

Introduction

The evolution of the human genome is one of the major interests in biological research. Population genetics has contributed a lot to our understanding of the evolutionary process (Crow and Kimura, 1970; Ewens, 2004; Hartl and Clark, 2007; Kimura and Ohta, 1971; Lynch and Walsh, 1998; Weir, 1996), especially through studies on gene genealogies using coalescence of individual sequences (Hudson, 1990; Rosenberg and Nordborg, 2002). The coalescent approach is a way of reconstructing ancestral history, in which mathematical tractability can provide easy incorporation of mutation and recombination in the reconstruction of ancestral histories (Nordborg, 2001). This method is mainly used to infer the population genetic parameters rather than infer the actual ancestral history of the genome due to its probabilistic properties. Also, because of its potential for capturing disease variants using population data, the fine-mapping method using a direct or indirect coalescent approach has been developed (Larribe *et al.*, 2002; Minichiello and Durbin, 2006; Molitor *et al.*, 2003a; Molitor *et al.*, 2003b; Molitor *et al.*, 2005; Rannala and Reeve, 2001; Rannala and Slatkin, 1998; Reeve and Rannala, 2002; Zollner and Pritchard, 2005; Zollner *et al.*, 2005).

In the inference of gene genealogies using coalescence, the incorporation of recombination in the coalescent process seems simple in that it only adds two ancestors to the graph. However, the actual construction of the ancestral recombination graph is extremely challenging (McVean and Cardin, 2005). It should be noted that frequent past recombination events over a large sequence imply multiple most recent common ancestors (MRCA), since recombination can add more ancestors than reduce ancestors by coalescence. In addition to the problems with incorporating recombination, the coalescent approach deals with individual sequences, which are different from the genotype data that we usually have. Therefore, in the coalescent approach, the stochastic inferences using genotype data involve additional assumptions or constraints (Zollner and Pritchard, 2005). This problem continues when using sequenced data because the positions of mutation sites are given.

In order to find a possible future remedy for the current approaches and obtain more descriptive analyses from actual ancestral graphs instead of probability distributions of graphs, an alternative way of constructing ancestral graphs that avoids the multiple MRCA, as well as unnatural constraints, is proposed in this study. Instead of constructing the genealogies by coalescence of each individual sequence in a backward direction, the focus is on the emerging order of variants in the ancestral history of genetic data. By concentrating on the variants themselves and constructing the graph in a forward direction, multiple MRCAs are avoidable, naturally

*Corresponding author: E-mail lypark@yonsei.ac.kr
Tel +82-2-2123-7615, Fax +82-2-313-8892
Accepted 16 February 2009

representing only one ancestral haplotype. The constructed ancestral graph exhibits only "observable" mutations and recombinations, thereby eliminating the unnaturally constrained mutation and unobserved recombination events. The proposed method is faithful to the observed data in constructing the past genealogy by reducing parameters that are unknown and not imperative.

In the proposed method, the estimation of allele age is critical for constructing the ancestral graph. Classically, allele ages are estimated from intra-allelic variability or allele frequency, considering population factors such as natural selection, genetic drift, mutation, and gene flow (Slatkin and Rannala, 2000). Since allele frequencies can be affected by unknown population factors, the exact measure of allele age is quite difficult to determine in most cases. The estimation of allele age from the linked markers can be distinguished by two major approaches (Rannala and Bertorelle, 2001), *i.e.*, a phylogenetic approach with a direct age inference from the root of the constructed tree and the population genetic approach that relies on models of demography, mutation, and recombination. The phylogenetic method does not consider recombination, but population genetic approaches may consider both mutation and recombination processes.

In population genetic approaches with recombination, the basic concept of using recombination in age estimation is the moment estimator of linkage disequilibrium (LD) decay (Slatkin and Rannala, 2000). For better estimates with multiple-linked marker loci, parametric statistical methods that incorporate demographic parameters have been developed with consideration of the gene tree (Rannala and Bertorelle, 2001). There has been an improvement in estimating allele ages using the approximation of coalescent approaches considering both allele frequency and LD (Rannala and Reeve, 2003; Slatkin, 2008). Relying on stochastic processes, these approaches are designed mainly for estimating the age of low-frequency disease mutations in conjunction with finding the locations of the disease mutations or for jointly estimating low-frequency allele age and selection intensity. Since the focus of this study is to derive the ancestral graph from the allele age of each variant, a new method for finding at least the relative allele ages of common variants is necessary.

To solve the problem as simply as possible, the demographic and population genetic parameters are considered to be minimal. Thus, the LD decay at linked markers is the only measurement in the age estimation of this study, which is partially similar to the methods of the moment estimator using LD (Slatkin and Rannala, 2000). In contrast with the moment estimator of LD, the recursive expression for LD decay is used for each age in consideration of the initial LD state, and a useful algorithmic procedure is developed for the mean allele age. Using the estimated ages, the ancestral graph can be derived. The basic idea for constructing the ancestral graph is to add new incoming alleles to the original haplotypes depending on the order of allele age. In this method, the initial LD state between the emerging variant and the existing variant(s) is critical in determining the node of the ancestral graph. To test this new method, a coalescent-based simulation sample and real data are examined in order to construct the ancestral graph.

Methods

Allele age estimation

For two biallelic single nucleotide polymorphisms (SNPs), there is a maximum of four possible haplotypes, and they decay together in a comprehensive manner as indicated in Equation (1). When the major alleles of each variant are "1" and the minor alleles are "2," the notations p_1, p_2, p_3, and p_4 indicate the frequencies of haplotypes 11, 12, 21, and 22, respectively. If R is the recombination fraction between two biallelic loci ($0<R\leq 0.5$), the haplotype frequencies ($p_{i,t}$) of the current generation are expressed as the haplotype frequencies at the previous time t-1 in the Wright-Fisher model with the assumptions of random mating and discrete non-overlapping generations. The decay generation, counted from Equation (1), can be used as the estimation of the relative allele age. This method can be applicable to any SNPs, but needs some corrections when applying to other types of polymorphisms such as short tandem repeats (STR).

$$p_{i,t}=p_{i,t-1}+\eta_i R(p_{2,t-1}p_{3,5-1}-p_{1,t-1}p_{4,t-1}) \text{ i=1, 2, 3, and 4 } \eta_1=\eta_4=1 \quad \eta_2=\eta_3=-1 \qquad (1)$$

LD between two variants decays after the emergence of the younger variant, which is indicated in Fig. 1 as dashed lines for examples. In Fig. 1, 11 → 21 involves mutation at the first locus (from allele 1 to 2), and this happened before the mutation at the second locus. 21 →22 in (a) and 11 → 12 in (b) involves mutation at the second locus. Since it is just an example figure, the emergence times of the younger variants are arbitrary for both (a) and (b) in Fig. 1. LD decay occurs in only one direction towards linkage equilibrium (LE) with no reversal. The recombination between two different haplotypes generates an existing haplotype (Fig. 1, 21 in (a) or 11 in (b)) and a new recombinant haplotype (Fig. 1, 12 in (a) or 22 in (b)). The LD decay state depends on the current frequencies of four possible haplotypes,

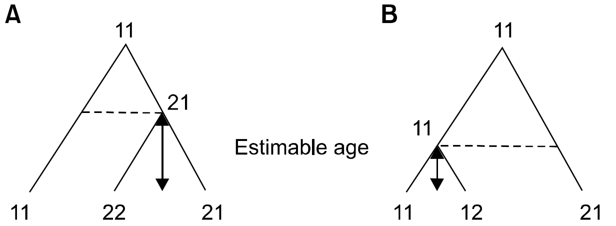

Fig. 1. Examples of ancestral history between two variants. 11→21 involves mutation at the first locus (from allele 1 to 2), and 21→22 in (A) and 11→12 in (B) involves mutation at the second locus. The time when mutation occurs is arbitrary in this figure.

which can be expressed by the major haplotype frequency when the allele frequencies are fixed. The frequency range of the major haplotype is limited, depending on the minor allele frequencies (Park, 2007). The current major haplotype frequency indicates one of the middle states between LE and complete LD. Since the allele age can be different depending on the initial state of LD ((a) or (b) in Fig. 1), determination of that state is important for allele age estimation. As the major haplotype frequency approaches the minimum, the LD decay state changes from the state (b) in Fig. 1 to the complete LE. In the opposite case, the initial state of LD decay changes from the state (a) in Fig. 1 to the complete LE.

The allele age can be calculated by subtracting the complete decay generation of the current LD state from the complete decay generation of the initial state. Instead of direct generation counts from the initial to current LD decay state, this method can provide a more accurate stopping point by avoiding pass or premature determination of the current decay state. Without drift, haplotype frequencies are changed only in the direction of LE until reaching LE. Using equation (1), the number of generations from the initial complete LD to the current state can be calculated for every two variants.

To reduce the usage of unknown parameters, the assumptions include fixed allele frequencies, fixed recombination rates, an infinite population size, no recurrent mutation, and no drift. Under these conditions, the LD decay between polymorphisms is dependent solely on the allele age. This method is basically similar to the methods of moment estimator in allele age estimation (Slatkin and Rannala, 2000), but it is much enhanced, based on the LD information from multiple linked variants and the usage of all observable haplotypes. The method serves as an efficient algorithmic procedure in consideration of the initial haplotype states. In case of LE in this method, the allele age is older than the generation that reaches linkage equilibrium, and complete LD means zero generation of the allele age. The recombination rate is assumed to be 1×10^{-8} per base pair (Strachan and Read, 2000) and is adjusted for each pair of variants, depending on the distance between variants.

The age obtained from the haplotypes of two variants is for only one of the variants, more precisely, the youngest allele among four possible alleles in two variants. In this method, the estimated age is actually for the allele with the lowest frequency due to the assumption of fixed allele frequencies and their initial states, which are derived from their haplotype frequencies. Consequently, variants with low frequencies have more data. The most frequent variant does not have any data remaining, and it is, consequently, the oldest one. If there are a total of k variants, there is a maximum of k-1 ages for a variant (maximum, k-1 ages, for the lowest frequency variant). For a variant's age, the mean can be the descent estimate of the variant age. To reduce the variance, the youngest one is removed from further age estimation of the other variants at each stage. The procedures are listed below:

(1) Calculate ages depending on their initial LD state from all possible LD combinations for k variants.

(2) Among ages of k-1 combinations for each variant, take the age into account only if the minor allele frequency of the target variant is less than the coupled variant and then find the mean for each variant.

(3) Find the youngest variant from the mean allele ages and remove all the data of other variants generated with the youngest variant. Return to the second step and repeat until the last variant remains.

As indicated, from the simple algorithmic procedure, the expected mean allele ages are assigned to each variant. For calculating the LD decay generation from the initial and current states, C++ was used for fast calculation. The stopping point for the age calculation is when the differences become lower than 10^{-5} between frequencies at time t and frequencies at complete LE. The algorithm for finding relative allele ages uses R.

Constructing the ancestral graph

The ancestral graph can be derived from the relative allele ages. This ancestral graph method is more oriented to the variant's origin than to individual sequences, different from the basic coalescent approach with mutation. In this method, the original haplotype, similar to MRCA, the root, consists of the original alleles of each variant. Basically, the role of relative allele age in constructing the ancestral graph is to order the emer-

Table 1. Order of age and ancestral state of the haplotypes from variants, A, B, and C. The ancestral states, (a) and (b), indicate the states of Fig. 1a and Fig. 1b

Variant	Age order*	With variant A	With variant B	With variant C
A	3	NA	a	a
B	1	-	NA	b
C	2	-	-	NA

*From the youngest.
NA, not applicable.

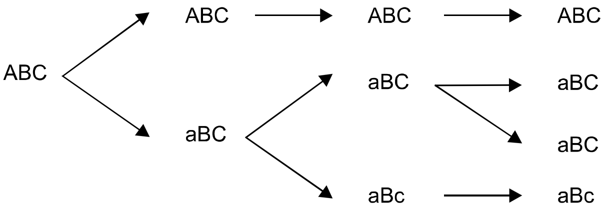

Fig. 2. Ancestral graph of three example variants using allele ages and LD decay statuses. "B" is the youngest and "A" is the oldest allele. "a", "b" and "c" indicate the mutated variants of "A", "B" and "C", respectively.

gence of variants and interpret proper spacing between nodes in the graph. The basic idea for constructing the ancestral graph is to add new mutations to the original haplotypes depending on the order and the initial state between variants. The original state between variants is very important in determining the original haplotype from which a new variant emerges. Therefore, the ancestral graph is obtained from the information of the allele ages and the initial LD state between variants.

A case of three variants, A, B, and C, is described here as an example. The order from the youngest to the oldest is B, C, and A. The LD decay status of each is determined as indicated in Table 1, in which the emergence status of mutation is indicated as (a) or (b). The status (a) and (b) indicate the statuses of Fig. 1a and Fig. 1b. Starting from the oldest variant, A, a mutation arises from the original haplotype ABC. At this time, the variants B or C have not emerged yet, and all B and C variants are the ancestral states "B" and "C" instead of the mutated ones "b" and "c." Next, the emergence of "c" allele is from the haplotype "aBC," since the initial LD decay status between A and C is (a). Therefore, there are three haplotypes, ABC, aBC, and aBc in this generation. For the youngest "b" allele, the LD decay status is (a) with A variant and (b) with C variant as shown in Table 1, which means that the b allele arises from the aBC haplotype, as shown in Fig. 2.

In a more complicated situation involving recombination, the target haplotype, in which a new mutation needs to arise, may not exist in this basic ancestral graph. Recombination between proper haplotypes can generate the target haplotype, and the recombinant haplotypes will remain after that. The time of recombination is between the time when the recombining haplotypes are generated and the time when the new variant comes out. The final ancestral graph derived from the allele age and the LD decay statuses of this example is indicated in Fig. 2. If there are haplotypes not generated from the ancestral graph, proper recombination events can be incorporated to generate the haplotypes.

A summary of the procedure for constructing the ancestral graph using R is presented below:

(1) Find the initial states between variants and the order of variants' ages.
(2) From the existing node(s), generate a new node of the oldest variant among the remaining variants, considering the initial states with the previously emerged variants. If necessary, generate a proper node by recombination and a new node from the newly generated proper node.
(3) Repeat step (2) until the final node is generated.
(4) Compare the generated haplotypes to the current existing haplotypes and incorporate recombination using most likely combinations of haplotypes depending on the haplotype frequencies and the number of recombination (*i.e.* single recombination is preferred). The timing of recombination events is right after the recombining haplotypes are generated.

This method is rather descriptive compared to the coalescent method. The most important difference is that the proposed method focuses on the variant itself, and the individual sequences are not considered in the basic framework. The underlying assumption is that no recurrent mutation occurs, implying that all the same alleles are considered to be identical by descent (IBD), which is reasonable in most cases considering estimated mutation rates and the theoretical result in population genetics (Crow and Kimura, 1970; Nachman and Crowell, 2000). One significant advantage of this method is the potential of the direct inference of ancestral mutation and recombination events. Further study would improve this method by estimating the accurate timing of recombination and the precise estimation of mean allele ages.

Sample description

To test the proposed ancestral graph method, a sample

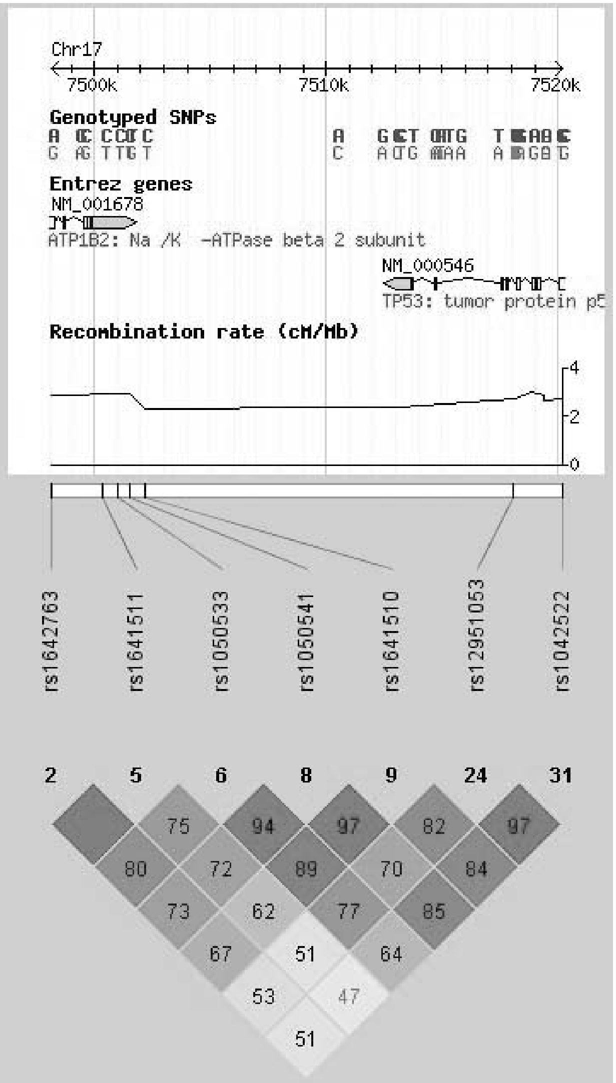

Fig. 3. Summary of the selected region (HapMap data of *TP53* region in Chromosome 17) for analysis from Haploview Version 4.0. The linkage disequilibrium (D') between selected SNPs is indicated in the bottom.

was generated using a program, ms, that uses the Wright-Fisher neutral model (Hudson, 2002). A total of 100 sequences were sampled for five segregating sites. The fraction of recombination based on a finite-sites recombination model (Hudson, 1983) was set to 100, based on the assumptions of a recombination probability of 10^{-8} per base pair, a population size of 2.5×10^6, and a total of 1,001 base pairs. The relative positions of variants were 0.3404, 0.4542, 0.6029, 0.6413, and 0.7169 for each variant.

To test real data, the genotype data from the HapMap project were used (2003; 2005). Among the publicly available genetic data, the region of tumor protein p53 (*TP53*) was selected considering the possibility that the genes involved in the fatal function of a life would have less selective or deleterious pressure on their common polymorphisms. Since allele age estimation method depends on recombination, the selected region is the region of recombination hotspots that have consistently high recombination rates, as determined by the HapMap project (Fig. 3) (2005). Among the accessible population groups, only Han Chinese in Beijing, China, (CHB) and Japanese in Tokyo, Japan, (JPT) are unrelated individuals. Since the current study deals only with unrelated individuals, CHB and JPT were selected for the study. The CHB and JPT populations, which show similar LD patterns, were combined for the analysis to obtain better estimates. To reduce sampling error, only minor allele frequencies of more than 0.1 were selected for the analysis, and seven single nucleotide polymorphisms (SNPs) in the region were finally selected (Table 2).

Results

Simulation-based sample

From a simulation-based sample for a test, the 100 sequences harboring five variants were generated as described in the sample description, and the ages and the

Table 2. Summary table of variants used in constructing ancestral graph from the real data of HapMap *TP53* region

Variant	No. in Fig. 3	Name	Position	HWE p-value	%Geno	MAF	Alleles*
1	2	rs1642763	7498144	0.4247	100	0.406	G:A
2	5	rs1641511	7500402	0.4267	98.9	0.421	T:C
3	6	rs1050533	7501019	0.7523	98.9	0.461	T:C
4	8	rs1050541	7501560	0.0903	100	0.378	T:G
5	9	rs1641510	7502221	0.0821	100	0.417	T:C
6	24	rs12951053	7518132	0.3879	100	0.361	A:C
7	31	rs1042522	7520197	0.7807	98.9	0.449	G:C

*First one is the major allele.
HWE, hardy-weinberg equilibrium; %Geno, percent of completed genotyping; MAF, minor allele frequency.

Table 3. The ages and ancestral states of variants generated by a simulation

Variant	MAF	Age order*	Mean age	With variant1	With variant2	With variant3	With variant4	With variant5
1	0.15	4	195342	NA	NA “B”	0 “B”	195342 “B”	NA “A”
2	0.13	3	145378	0 “B”	NA	213018 “A”	290755 “A”	NA “B”
3	0.15	1	0	0 “B”	NA “A”	NA	0 “A”	NA “B”
4	0.45	5	NA	NA “B”	NA “A”	NA “A”	NA	NA “B”
5	0.09	2	135806	407418 “A”	0 “B”	0 “B”	0 “B”	NA

*The numbers from the youngest.

“ ”, the initial haplotype state indicated in Fig. 1; Number above “ ”, allele ages from LD; MAF, minor allele frequency; NA, not applicable.

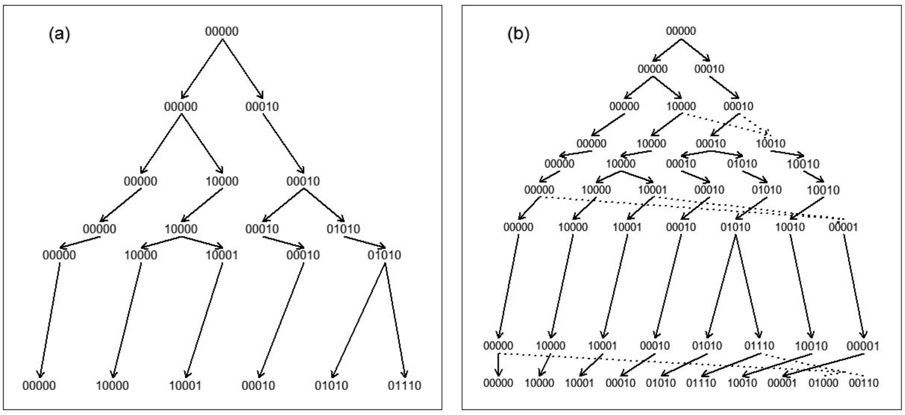

Fig. 4. Ancestral graphs from the simulated haplotype data. (a) basic ancestral graph (b) ancestral recombination graph (recombination is indicated as dashed lines). The original allele is indicated as “0”, and the mutated allele is indicated as “1”. The only newly emerged recombinant is indicated in (b).

ancestral states of five variants were determined (Table 3). As described in the methods section, only the ages of the variants that have lower frequencies than the coupled variants through LD are considered and are indicated in Table 3. The mean age is not always the mean of all indicated ages in the Table since the ages coupled with the youngest variants are eliminated for the mean ages of other variants after each step to reduce variance. In this method, the complete LD always results in zero generation because allele ages depend only on the LD. The mean age of the oldest variant cannot be found since there are no data left for estimating its allele age. The order of allele age is obtained from the mean allele ages of variants. From this information, the ancestral graph for the observed five variants is drawn as shown in Fig. 4a, and the ancestral recombination graph is drawn as shown in Fig. 4b. Starting from the common ancestral haplotype 00000, the occurrence of each variant generates a new node. For examining the haplotypes in the current generation, the individual haplotypes from the sample sequences are summarized in Table 4.

Table 4. Individual haplotypes and their numbers from the simulation data

Index	Number	Haplotype
1	9	10000
2	34	00000
3	10	01110
4	27	00010
5	3	10010
6	3	01000
7	5	00110
8	6	00001
9	3	10001
Total	100	-

Underline, possible recombinants.

The detailed interpretation of the ancestral graph derived from this test sample is described below and includes the putative recombination events. Starting from the ancestral graph without recombination, among the current haplotypes at the final nodes in the ancestral graph, only haplotype 01010 disappears in the current sequences, while the rest of the haplotypes remain. Haplotype 01010 has the same age as variant 2, which has a mean age of 145378 generations. However, the haplotype should exist until the appearance of variant 3, the mean age of which is 0 generations. Although this is a simulated sample with a small sample size of 100 sequences using unrealistic assumptions due to the base model of the simulation, it can be predicted that haplotype 01010 disappeared quite recently in the simulation data based on the current study.

Even though this test is based merely on simulated data, there are four more haplotypes that were not seen in the ancestral graph without recombination: 10010, 01000, 00110, and 00001. These are recombinants from the existing haplotypes, showing relatively low frequencies, as indicated in Table 4. Ignoring double recombination and focusing on the haplotypes in the graph, the probable recombination pairs are summarized in Table 5. Each recombination event happened in a boundary of limiting generations depending on the ages of the haplotypes. Here, the recombination is incorporated into the ancestral graph at the right after the recombining haplotype is generated (Fig. 4b). Among the possible recombination events in Table 5, the events involving the submerged haplotype and the haplotypes with lower frequencies are excluded for ARG as in Fig. 4b, since the other recombination events are enough for generating the target haplotypes.

Table 5. Probable combination of recombination from the simulation data

Recombination pair		Recombinants	
10000	00010	10010	00000
10001	00010	10010	00001
10000	01010*	10010	01000
01110	00000	01000	00110
01110	00010	01010*	00110
10001	00000	00001	10000

*Submerged haplotype. Underline, newly generated recombinant haplotype that was not shown in Fig. 4a.

Real data

For testing real data, the genotype data of protein p53 (*TP53*) from the HapMap project were used (2003; 2005). A detailed description of the selection criteria is in the methods section. From the seven SNPs selected (Table 2), the calculated allele ages are summarized in Table 6. As described in the methods section, the most frequent variant, variant 3, is the oldest. It happens here that the least frequent variant, variant 6, is the youngest. Interestingly, the minor allele frequency range of all seven variants is small in this data set. The initial states between all variants are the state "A," which means the new allele comes from the haplotype harboring the younger allele (minor allele in this study) of the existing older variant. Therefore, the ancestral graph can be drawn as shown in Fig. 5, Table 6 shows that, except for the two oldest variants, all variants have ages be-

Table 6. The ages and ancestral states of variants from the HapMap data in *TP53* region

	MAF	Age order*	Age	V1	V2	V3	V4	V5	V6	V7
1	0.406	3	5746	NA	0 "A"	7460 "A"	NA "A"	9779 "A"	NA "A"	2999 "A"
2	0.421	6	45124	NA "A"	NA	45124 "A"	NA "A"	NA "A"	NA "A"	3782 "A"
3	0.461	7	NA	NA "A"	NA "A"	NA	NA "A"	NA "A"	NA "A"	NA "A"
4	0.378	4	14289	9045 "A"	27701 "A"	11163 "A"	NA	4003 "A"	NA "A"	839 "A"
5	0.417	5	17292	NA "A"	25459 "A"	9125 "A"	NA "A"	NA	NA "A"	920 "A"
6	0.361	1	2194	3174 "A"	3748 "A"	1504 "A"	2072 "A"	1177 "A"	NA	1489 "A"
7	0.449	2	2295	NA "A"	NA "A"	2295 "A"	NA "A"	NA "A"	NA "A"	NA

*From the youngest.
" ", the initial haplotype state indicated in Fig. 1; Number above " ", Allele ages from LD; MAF, minor allele frequency; NA, not applicable.

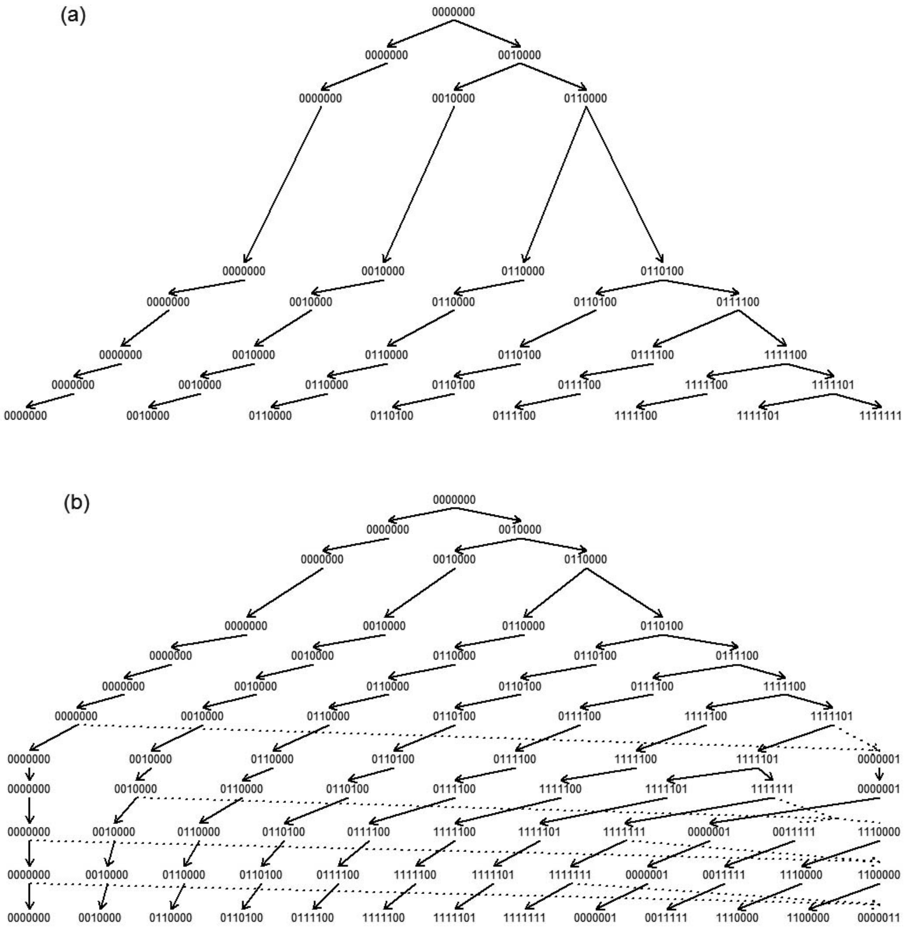

Fig. 5. Ancestral graphs from the real data. (a) basic ancestral graph (b) ancestral recombination graph (recombination is indicated as dashed lines). The original allele is indicated as "0", and the mutated allele is indicated as "1". The only newly emerged recombinant is indicated in (b).

tween 2194 to 17292 generations, a relatively short interval compared with the emergence time of 45124 generations for the second-oldest variant (variant 2). This is consistent with the short range of minor allele frequencies among the variants.

From the basic ancestral graph generated from the order of the allele ages, there are eight haplotypes (number of segregating sites+1), as shown in Fig. 5a. For comparison of the haplotypes in the graphs with the haplotypes estimated directly from the genotype data using existing methods (Table 7), the two most popular methods were used to estimate haplotypes and their frequencies. One is the accelerated EM algorithm similar to the partition/ligation method (Qin *et al.*, 2002), which is implemented in Haploview Version 4 (Barrett *et al.*, 2005). Another is the coalescent-based haplotype estimation using PHASE Version 2.1.1 (Stephens and Scheet, 2005; Stephens *et al.*, 2001). In Table 7, the probable original allele (major allele) is indicated as "0," and the changed allele is indicated as "1." Also, in representing haplotypes with original notation, the numbers "1, 2, 3, and 4" indicate "A, C, G, and T," respectively, as notated in Haploview (Table 7).

As summarized in Table 7, the haplotypes and their frequencies estimated from these two methods are completely different, except for a haplotype (1110111) and the two most common haplotypes (0000000 and 1111111), which are the original haplotype (0000000) and the haplotype that consists of altered alleles (1111111). Even the frequencies obtained by the two methods for those haplotypes are very different. In many cases, however, the inferred haplotypes are not very different between the methods. However, in this case, minor alleles of all the variants are linked together, comprising the two most common haplotypes, 0000000 and 1111111, which makes the inference of the rest of the haplotypes more difficult. The rest of the haplotypes obtained using the EM algorithm show minor allele frequencies of less than 0.05, but this is not the case using the coalescent method.

Interestingly, the method proposed in this study shows more similarity to the haplotype estimates using the EM algorithm. Haplotypes 1111111 and 1111101 are seen in haplotypes with frequencies higher than 0.01, as

Table 7. Comparison of haplotype frequencies (real data) from haploview and phase, and the existence in the constructed ancestral graph

Index	Haploview (EM algorithm)				PHASE (Coalescence)		
	Original	Simplified	Frequency	Fig. 5*	Haplotype	Frequency	Fig. 5*
1	3444413	0000000	0.419	Y	0000000	0.282	Y
2	1223222	1111111	0.235	Y	1111111	0.162	Y
3	3423222	0011111	0.048		0000100	0.154	
4	1223212	1111101	0.046	Y	1110111	0.101	
5	1224413	1110000	0.035		1010000	0.047	
6	1244413	1100000	0.031		0100111	0.035	
7	3444412	0000001	0.028		0000110	0.032	
8	3444422	0000011	0.025		1110101	0.030	
9	1223213	1111100	0.019	Y	0000010	0.029	
10	3424413	0010000	0.018	Y	1110110	0.027	
11	3424222	0010111	0.012		1010111	0.020	
12	3244413	0100000	0.011		0000111	0.011	
13	1224222	1110111	0.010		0100100	0.010	

*Y if the target haplotype exists in Fig. 5a. For original haplotypes, "1, 2, 3, and 4" denote "A, C, G, and T," respectively.

Table 8. Probable combination of recombination pairs and recombinants for top 10 frequent haplotypes estimated by EM algorithm (real data).

Recombination pair		Recombinants	
0000000	1111111	0011111 (3)	1100000 (6)
0010000	1111111	0011111 (3)	1110000 (5)
0000000	1111101	0000001 (7)	1111100 (9)
0000000	1111111	0000011 (8)	1111100 (9)

*( ), The number inside indicates the index number of Table 7.

estimated by EM (Table 7). Depending on the constructed ancestral graph in Fig. 5a, these haplotypes are generated most recently compared to other haplotypes in Fig. 5a, thereby indicating an increased probability of observing these two haplotypes in the current generation. Appending these two haplotypes, haplotype 1111100 is the third recent haplotype among the eight haplotypes in Fig. 5a, and the haplotype also can be seen among haplotypes with frequencies higher than 0.01 using EM inference (Table 7). The rest of the haplotypes in the EM estimates are expected to come from recombination events. The reason for differences with the coalescent-based method is scrutinized in the discussion section.

Similar to the simulation data, the probable combination of recombination for the top 10 frequent haplotypes as estimated by the EM inference is represented in Table 8. Among the haplotypes in Table 7, those not indicated in the basic ancestral graph (Fig. 5a) are only considered as recombinants, which are the haplotypes with index numbers, 3, 5, 6, 7, and 8 in Table 7. Primarily, single recombination is considered. As indicated in Table 8, the haplotypes that are not observed in Fig. 5a can now be generated by each single recombination event from the existing haplotypes of the basic ancestral graph, Fig. 5a. Interestingly, the third most frequent haplotype, 0011111, can be generated by recombination of either haplotype pairs 0000000 and 1111111, or 0010000 and 1111111. In Table 8, the four recombination pairs include three haplotypes of each of the most common haplotypes, 0000000 and 1111111, which are very likely to be involved in recombination events due to their high frequency. Taking the haplotype frequencies into account, the most likely recombination events in Table 8 are incorporated into the graph in order (Fig. 5b).

Discussion

A novel methodology for constructing an ancestral graph is proposed in this study. The proposed method is based on the variants themselves rather than individual sequences, so the method produces results that are based on actual genetic data rather than models. Although more developments are necessary, the main advantages of this method are that (1) it is computationally favorable, (2) it facilitates descriptive interpretation of the ancestral graph, and (3) it provides the easier and tangible incorporation of recombination into the graph. It is noteworthy that the proposed method can provide a descriptive ancestral graph rather than a probability distribution of all possible graphs as coalescent-based methods do. Therefore, the current method is designed mainly for obtaining the actual ancestral graph rather than inferring population genetic parameters that is the main purpose in most coalescent-based studies. At the current stage, the method of allele age estimation is just aimed at finding the relative allele ages of variants. For accurately constructing the ancestral graph, improvements are necessary for both the accurate allele age estimation and the accurate incorporation of recombination.

Problems in estimating allele age in this study can be mitigated if the effective population size is fairly large. When the effective population size is large, the allele frequency of a newly emerged mutation depends largely on the number of generations (Crow and Kimura, 1970; Kimura and Ohta, 1971). Therefore, it is more likely that the less frequent variant is younger than the more frequent one. Another problem with this method is the assumption of fixed allele frequencies during LD decay. In the binomial random process, the expected current allele frequency is the initial allele frequency due to its Markovian property. However, in the actual situation, it is more likely that the variant came from a single mutation and increased its frequency through random drift. Further advancements of allele age estimations can be achieved by considering the stochastic process of haplotypes in a population. Since the consideration of the stochastic process is only for allele age estimation, the descriptive nature of this ancestral graph method will remain. The actual incorporation of recombination in the ancestral graph can also be improved by further development of the method.

As shown in the results section, the haplotypes from the proposed method using real data explain the haplotypes estimated by the EM algorithm better than the coalescent method. Differences from the haplotypes estimated by the coalescent-based method may occur during the inference of genealogical histories by coalescence. As indicated previously, the underlying assumptions of the coalescent theory can influence the resulting haplotypes. Since the coalescent method is based exclusively on population demographic history, haplotype inference might be deviated from the actual data while running the haplotype inference. Since the EM method depends heavily on the data, the estimated haplotypes are statistically reasonable choices that can be observed from the given data. Therefore, it is encouraging that the proposed method represents haplotypes that are more similar to those based on the EM estimates than the estimates from the coalescent-based simulation methods. The methodology in the current study does not depend heavily on the population model or the parameters. Therefore, it is more likely that the method in this study might be influenced by unknown population factors to a lesser extent than the previous coalescent-based methods.

Regarding the population differences in the *TP53* gene region, there are clear differences in the LD patterns among different population groups. First, the allele frequencies are strikingly different. Some variants are frequent only in one population, and minor alleles of several variants are changed to major alleles in another population. Second, the linkage disequilibrium patterns are very different among ethnic groups. The samples from the Yoruba in Ibadan, Nigeria, (YRI) and from Utah, USA, (CEU) show more similarity in both allele frequency and LD pattern than the samples from Han Chinese in Beijing, China, (CHB) and Japanese in Tokyo, Japan (JPT). It seems clear that much more similarities in the polymorphic pattern are shown within Asian populations compared to other populations (Kim *et al*., 2008; Lee *et al*., 2008). The LD pattern of the combined sample of CHB and JPT shows much more linkage disequilibrium between variants than the other two populations.

It seems that the high recombination fraction in the *TP53* gene region from the HapMap project might come from the population effect of CEU and YRI, since the estimated recombination rates were the averages for all the populations (2005). Because only CHB and JPT populations are used as indicated in the methods section, the default recombination rate, 10^{-8}, described in the methods section (Strachan and Read, 2000), was used instead of 2×10^{-8}, as indicated in the HapMap data for the allele age estimation (Fig. 3). It should be noted that the proposed method is very sensitive to haplotype frequencies. Therefore, considering the population differences as mentioned earlier, caution would be necessary when interpreting the results from different population groups.

A novel ancestral graph method is proposed that focuses on variants rather than individual sequences. This deterministic method is less affected by past demographic histories and population genetic parameters than previous methods, and it presents a simpler way of constructing ancestral graphs, focusing only on the allele frequencies of variants and LD of the genotyped variants. By excluding unnecessary population genetic parameters, the method can provide a more practical interpretation of the human genome, as described previously. More importantly, the proposed method seems to represent both real and simulation data quite well. The results of the real data allows for a better fit with the haplotypes estimated using the EM algorithm rather than the coalescent-based simulated method.

The proposed method can incorporate the observed ancestral recombination event without the concern of unobtainable MRCA. Due to the potential for localizing the disease variant based on genealogical history, the coalescent approach has attracted attention (Rosenberg and Nordborg, 2002). However, as indicated previously, the inference of the ancestral graph with recombination using the coalescent method is computationally challenging (McVean and Cardin, 2005). Although there are good approximations on coalescence with recombination for mapping disease variants (Larribe *et al.*, 2002; Minichiello and Durbin, 2006; Molitor *et al.*, 2003a; Molitor *et al.*, 2003b; Molitor *et al.*, 2005; Rannala and Reeve, 2001; Rannala and Slatkin, 1998; Reeve and Rannala, 2002), a new method is necessary to infer the actual genealogical history. The proposed approach for constructing ancestral graphs could be a possible future remedy for the previous coalescent approaches in terms of computational efficiency and theoretical settlement of constraints in mutation and problems in incorporating recombination events.

Acknowledgements

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-532-C00017). The key calculations were performed using the supercomputing resource at the Korea Institute of Science and Technology Information (KISTI). The author thanks referees for helpful comments.

References

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263-265.

Crow, J.F., and Kimura, M. (1970). *An Introduction to Population Genetics Theory.* (New York: Harper & Row, Publishers).

Ewens, W.J. (2004). *Mathematical Population Genetics.* New York: Springer.

Hartl, D.L., and Clark, A.G. (2007). *Principles of Population Genetics.* (Sunderland: Sinauer Associates, Inc.).

Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23, 183-201.

Hudson, R.R. (1990). *Gene Genealogies and the Coalescent Process. Oxford Surveys in Evolutionary Biology.* (New York: Oxford University Press).

Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337-338.

Kim, Y.U., Kim, S.H., Jin, H., Park, Y.K., Ji, M.H., and Kim, Y.J. (2008). The Korean HapMap Project Website. *Genomics & Informatics* 6, 91-94.

Kimura, M., and Ohta, T. (1971). *Theoretical Aspects of Population Genetics.* (Princeton: Princeton University Press).

Larribe, F., Lessard, S., and Schork, N.J. (2002). Gene mapping via the ancestral recombination graph. *Theor. Popul. Biol.* 62, 215-229.

Lee, J.E., Jang, H.Y., Kim, S., Yoo, Y.K., Hwang, J.J., Jun, H.J., *et al.* (2008). Chromosome 22 LD map comparison between Korean and other populations. *Genomics & Informatics* 6, 18-28.

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* (Sunderland: Sinauer Associates, Inc.)

McVean, G.A., and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360, 1387-1393.

Minichiello, M.J., and Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* 79, 910-922.

Molitor, J., Marjoram, P., and Thomas, D. (2003a). Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. *Genet. Epidemiol.* 25, 95-105.

Molitor, J., Marjoram, P., and Thomas, D. (2003b). Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am. J. Hum. Genet.* 73, 1368-1384.

Molitor, J., Zhao, K., and Marjoram, P. (2005). Fine mapping - 19th century style. *BMC Genet.* 6 Suppl 1, S63.

Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297-304.

Nordborg, M. (2001). Coalescent theory. In: *Handbook of Statistical Genetics,* D.J. Balding, M. Bishop, C. Cannings ed. (New York: Wiley), pp. 179-212.

Park, L. (2007). Controlling linkage disequilibrium in association tests: revisiting APOE association in Alzheimer's disease. *Genomics & Informatics* 5, 61-67.

Qin, Z.S., Niu, T., and Liu, J.S. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 71, 1242-1247.

Rannala, B., and Bertorelle, G. (2001). Using linked markers to infer the age of a mutation. *Hum. Mutat.* 18, 87-100.

Rannala, B., and Reeve, J.P. (2001). High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am. J. Hum. Genet.* 69, 159-178.

Rannala, B., and Reeve, J.P. (2003). Joint Bayesian estimation of mutation location and age using linkage disequilibrium. *Pac. Symp. Biocomput.* 526-534.

Rannala, B., and Slatkin, M. (1998). Likelihood analysis of disequilibrium mapping, and related problems. *Am. J. Hum. Genet.* 62, 459-473.

Reeve, J.P., and Rannala, B. (2002). DMLE+: bayesian linkage disequilibrium gene mapping. *Bioinformatics* 18, 894-895.

Rosenberg, N.A., and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3, 380-390.

Slatkin, M. (2008). A Bayesian method for jointly estimating allele age and selection intensity. *Genet. Res.* 90, 129-137.

Slatkin, M., and Rannala, B. (2000). Estimating allele age. *Annu. Rev. Genomics. Hum. Genet.* 1, 225-249.

Stephens, M., and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* 76, 449-462.

Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978-989.

Strachan, T., and Read, A.P. (2000). *Human Molecular Genetics.* John Wiley & Sons (Asia) Pte Ltd.

The International HapMap Consortium (2003). The International HapMap Project. *Nature* 426, 789-796.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* 437, 1299-1320.

Weir, B.S. (1996). *Genetic Data Analysis II (2nd ed.).* Sunderland: Sinauer Associates, Inc.

Zollner, S., and Pritchard, J.K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* 169, 1071-1092.

Zollner, S., Wen, X., and Pritchard, J.K. (2005). Association mapping and fine mapping with TreeLD. *Bioinformatics* 21, 3168-3170.