Published in final edited form as:
Commun Stat Theory Methods. 2009 ; 38(18): 3445–3460. doi:10.1080/03610920802696588.

Modeling Informatively Missing Genotypes in Haplotype Analysis

Nianjun Liu[1], **Richard Bucala**[2,3], and **Hongyu Zhao**[3,4,*]

[1]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL

[2] Departments of Internal Medicine and Pathology, Yale University School of Medicine, New Haven, CT

[3]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT

[4]Department of Genetics, Yale University School of Medicine, New Haven, CT

Abstract

It is common to have missing genotypes in practical genetic studies. The majority of the existing statistical methods, including those on haplotype analysis, assume that genotypes are missing at random—that is, at a given marker, different genotypes and different alleles are missing with the same probability. In our previous work, we have demonstrated that the violation of this assumption may lead to serious bias in haplotype frequency estimates and haplotype association analysis. We have proposed a general missing data model to simultaneously characterize missing data patterns across a set of two or more biallelic markers. We have proved that haplotype frequencies and missing data probabilities are identifiable if and only if there is linkage disequilibrium between these markers under the general missing data model. In this study, we extend our work to multi-allelic markers and observe a similar finding. Simulation studies on the analysis of haplotypes consisting of two markers illustrate that our proposed model can reduce the bias for haplotype frequency estimates due to incorrect assumptions on the missing data mechanism. Finally, we illustrate the utilities of our method through its application to a real data set from a study of scleroderma.

Keywords

haplotype; genotype; missing at random; informatively missing; linkage disequilibrium; haplotype frequency; single nucleotide polymorphisms; multi-allelic marker; EM algorithm

1. Introduction

Haplotypes (the combination of alleles on the same chromosome that were inherited as a unit) are widely used in genetic studies (Schaid, 2004a), such as in population genetics and disease gene mapping. In population genetics, haplotypes have been shown to be more informative than single markers, and they have been used to study migration and immigration rates, genetic demography, and evolutionary history (Harding et al., 1997; Kidd et al., 2000; Tishkoff et al., 2000; Zhao et al., 2007). In disease gene mapping, studies have shown that methods based on haplotypes may be more powerful than those based on single markers (Akey, Jin, & Xiong, 2001; Botstein & Risch, 2003; Fallin et al., 2001; Morris & Kaplan, 2002; Schaid, 2004a; Schaid, Rowland, Tines, Jacobson, & Poland, 2002; Yu & Schaid, 2007b; H. Zhao et al., 2000). One reason for their importance is that haplotypes play a critical role in our

[*]Correspondence to: Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520-8034. hongyu.zhao@yale.edu. Tel: (203) 785-6271. Fax: (203) 785-6912.

understanding of linkage disequilibrium (LD, the non-random co-occurrence of alleles at two or more loci), which is essential in genetic studies. Usually, haplotypes are not directly obtained from experimentation because of the high cost of conducting such experiments (Excoffier, Laval, & Balding, 2003; Li, Khalid, Carlson, & Zhao, 2003; Tregouet, Escolano, Tiret, Mallet, & Golmard, 2004). Genotyping family members is an alternative way to obtain haplotypes, but there are practical limitations to this approach (Li et al., 2003; Wijsman, 1987; Zhang, Sun, & Zhao, 2005). As a result, numerous analytical methods have been proposed to infer haplotypes (Clark, 1990; Excoffier et al., 2003; Excoffier & Slatkin, 1995; Fallin & Schork, 2000; Hawley & Kidd, 1995; Li et al., 2003; Long, Williams, & Urbanek, 1995; Niu, Qin, Xu, & Liu, 2002; Stephens & Scheet, 2005; Stephens, Smith, & Donnelly, 2001; Zhang, Sheng, Morabia, & Gilliam, 2003). Once haplotype frequency estimates are obtained, they can be used in haplotype-based association analysis (Chiano & Clayton, 1998; Fallin et al., 2001; Schaid, 2004b; Schaid et al., 2002; Zaykin et al., 2002a). Statistically more powerful methods also have been proposed to simultaneously estimate haplotype frequencies and their effects (Epstein & Satten, 2003; Lin, Zeng, & Millikan, 2005; Zhao, Li, & Khalid, 2003).

Even with the advancement of modern biotechnologies, data with missing genotypes are still common in genetic studies. In addition to the possibility of missing data being caused by the equipment itself, such as any damage to or loss of performance by probes in multiplexed genotyping platforms, there are other circumstances that contribute to absent data, such as variation in DNA quality or molecular effects, experimental techniques used (including the genotype calling algorithms), and the ways in which studies are conducted, which can cause some individuals (e.g., cases versus controls in case-control studies) and some sites to have more or less than a reasonable share of missing data (Di et al., 2005; Hao & Cawley, 2007; Liu, Beerman, Lifton, & Zhao, 2006; Nicolae, Wu, Miyake, & Cox, 2006; Stephens & Scheet, 2005).

Existing methods handle missing genotype data in mainly two ways. The first way is to obtain complete data by either eliminating or imputing missing data prior to analysis (Hastie, Tibshirani, & Friedman, 2001; Yu & Schaid, 2007a; Zhang et al., 2003). The second way is to handle missing genotypes in the model, usually by assuming either explicitly or implicitly that missing is “at random” (i.e., at a given marker, different genotypes and different alleles are missing with the same probability and that the “missingness” at different markers is independent), both in haplotype frequency estimation (Excoffier & Slatkin, 1995; Li et al., 2003; Stephens et al., 2001; Tregouet et al., 2004) and in haplotype association analysis (regardless of the frequencies of genotypes or alleles at the same marker, regardless of the frequencies of haplotypes under consideration, and regardless of phenotype—i.e., disease status) (Epstein & Satten, 2003; Tregouet et al., 2004). This missing data mechanism assumption is likely to over-simplify the reality. In practice, even very carefully designed studies, such as HapMap and the Wellcome Trust Case Control Consortium, have informatively missing genotypes (Hao & Cawley, 2007; The Wellcome Trust Case Control Consortium, 2007).

In our previous work (Liu et al., 2006), we showed that haplotype frequency estimates can be biased when using methods that assume missing at random if this assumption is violated. Similarly, haplotype association analysis may be biased as well, inducing both false-positive and false-negative evidence of association. To overcome this problem, we proposed a general model to characterize the missing data mechanism across a set of two or more biallelic markers simultaneously. Under our model for missing data, we proved that haplotype frequencies and missing data probabilities are identifiable if and only if there is LD between these markers. A simulation study showed that our proposed model can reduce biases caused by incorrectly assuming missing at random, thus providing more accurate statistical inference.

In this article, we extend our previous work to markers with multiple alleles. We investigate the relationship between the identifiability of our model parameters and LD between the markers. In addition, we use simulations based on two markers to illustrate that our proposed model can reduce biases induced by inappropriate handling of missing genotype data, thus providing more accurate statistical inference. Our method is then applied to a real data set from a study of scleroderma to demonstrate its utility in practice.

2. Proposed Missing Data Model

2.1 Notation and Assumptions

Let $G = g = \{g_1, g_2, ..., g_L\}$ denote a subject's genotype at L markers, g_i denote the genotype at marker *i*, and $H = (h, h')$ denote haplotype pair, *h* and *h'*, carried by this subject, where (*h, h'*) is considered as ordered to avoid a factor of 2 in the likelihood functions when the two haplotypes are different. If L = 1, *G* and *H* are both genotypes at the single marker. However, because (*h, h'*) is ordered, we distinguish the parental origin of the two alleles at the single marker when we use *H*. We let *S(g)* denote the set of haplotype pairs $\{H = (h, h')\}$ that are consistent with $G = g$, p_h denote the frequency of haplotype *h* in the study population, n_g denote the number of individuals with genotype *g*, and *n* denote the sample size. For simplicity, we consider only two markers in the following analysis, and the extension to multiple markers is straightforward. Denote the two markers as A and B, and assume that these two markers have M and N alleles ($M,N \geq 2$), respectively. Let $\{A_1, A_2, ..., A_M\}$ be the M alleles of marker A and $\{B_1, B_2, ..., B_N\}$ be the N alleles of marker B. Let $p_{h_{A_rB_s}}$ denote the frequency of a haplotype consisting of two alleles, A_r and B_s, at the two markers A and B, respectively, and let p_{A_r} and p_{B_s} denote the two allele frequencies. We use α and γ to denote missing probabilities at markers A and B, respectively, and we assume that missingness is independent between markers and that there is Hardy-Weinberg equilibrium (HWE) for the two markers in the general population.

2.2 Missing Data Model

We have proposed a missing data model for biallelic markers such as SNPs (Liu et al., 2006). For one SNP with two alleles, A and B, Table 1 in Liu et al. (2006) shows the genotype penetrances—i.e., the conditional probability of observing one genotype given the true genotype. We define the probabilities related to missingness as follows.

$$\begin{aligned}
\alpha_1 &= \Pr(G = A_1?|H = (A_1, A_1)) \\
\alpha_2 &= \Pr(G = ??|H = (A_1, A_1)) \\
\alpha_3 &= \Pr(G = A_1?|H = (A_1, A_2)) \\
\alpha_4 &= \Pr(G = ??|H = (A_1, A_2)) \\
\alpha_5 &= \Pr(G = A_2?|H = (A_1, A_2)) \\
\alpha_6 &= \Pr(G = ??|H = (A_2, A_2)) \\
\alpha_7 &= \Pr(G = A_2?|H = (A_2, A_2)),
\end{aligned}$$

where "?" stands for a missing allele and M = 2—i.e., marker A has two alleles, A_1 and A_2. This missing data model is very general, and the missing at random model corresponds to $\alpha_1 = \alpha_3 = \alpha_5 = \alpha_1, \alpha_2 = \alpha_4 = \alpha_6 = \alpha_1^2$. Note that we assume that it is possible to observe partial genotypes—e.g., $A_1?$. We showed that the parameters of the above model are not identifiable if only one SNP is studied, and we further showed that for two SNPs, the model parameters (i.e., haplotype frequencies and missing data probabilities) are identifiable if and only if there is LD between the two markers (Liu et al., 2006).

In practice, most of the time we do not have partial calls (e.g., only one of the two alleles is missing at one marker), and sometimes even when there are partial calls investigators simply do not report them, but only report complete calls (Liu et al., 2006). Therefore, in this study we only consider complete calls. For a marker with K alleles, there are $\left(\begin{array}{c}K\\2\end{array}\right)+K$ possible genotypes (without considering missing genotypes). We define the probabilities (i.e., the penetrances) related to missingness as follows for a marker with three alleles denoted as A$_1$, A$_2$, and A$_3$:

$$\begin{array}{ll}\alpha_{1,1} & =\Pr\left(G=??|H=(A_1,A_1)\right)\\ \alpha_{1,2} & =\Pr\left(G=??|H=(A_1,A_2)\right)\\ \alpha_{1,3} & =\Pr\left(G=??|H=(A_1,A_3)\right)\\ \alpha_{2,2} & =\Pr\left(G=??|H=(A_2,A_2)\right)\\ \alpha_{2,3} & =\Pr\left(G=??|H=(A_2,A_3)\right)\\ \alpha_{3,3} & =\Pr\left(G=??|H=(A_3,A_3)\right).\end{array}$$

Table 1 shows the genotype penetrances for the model for this marker.

For a marker with K alleles, there are $\left(\begin{array}{c}K\\2\end{array}\right)+K$ degrees of freedom from the data if missing genotypes are observed. There are $\left(\begin{array}{c}K\\2\end{array}\right)+K$ parameters for missing probabilities and (K – 1) parameters for allele frequencies. The number of parameters exceeds the number of degrees of freedom, so under the above model the parameters are not identifiable if one marker is considered. If there are two markers, we have the following proposition, which can be viewed as a generalization of our previous finding for two biallelic markers.

Proposition: Under the above model with two markers, the model parameters (i.e., haplotype frequencies and missing data probabilities) are identifiable if and only if there is LD between the two markers.

Proof: Assume that we have two markers, A and B, under study, with the notations defined above. We have proved the proposition for two biallelic markers in our previous work (Liu et al., 2006). To prove the current proposition, we organize the proof into three steps. In step 1, we consider the simplest case, M = 3 and N = 2. In step 2, we generalize the simplest case to the case in which M >1 and N = 2. In step 3, we consider the general case in which M and N are arbitrary integers with M > 1 and N > 1. The amount of LD between alleles A$_r$ and B$_s$ can be measured by $D_{rs}=h_{A_rB_s}-p_{A_r}p_{B_s}$ (Kalinowski & Hedrick, 2001; Nothnagel, Furst, & Rohde, 2002). It is easy to see that for two bi-allelic markers the absolute values of the four D$_{rs}$'s are equal. D_{rs} = 0 (r = 1,---,M and s = 1,---,N) means that there is no LD—i.e., linkage equilibrium (LE)—between two markers A and B. Without loss of generality, we assume that all of the allele frequencies are strictly between 0 and 1.

Step 1: We focus on alleles A$_1$ and B$_1$ first. We can “pool” alleles {A$_2$, A$_3$} at marker A and denote it as A$_{-1}$, and we can think of A$_{-1}$ as a pseudo-allele. Now, we have two biallelic markers with alleles {A$_1$, A$_{-1}$} and {B$_1$, B$_2$}, respectively. The genotype penetrances for marker A are shown in Table 2, where

$$\begin{array}{ll}\beta_{1,-1} & =\Pr\left(G=??|H=(A_1,A_{-1})\right)=\Pr\left(G=??|H=(A_1,A_2)\quad \text{or}\quad H=(A_1,A_3)\right)\\ \beta_{-1,-1} & =\Pr\left(G=??|H=(A_{-1},A_{-1})\right)=\Pr\left(G=??|H=(A_2,A_2)\quad \text{or}\quad H=(A_2,A_3)\quad \text{or}\quad H=(A_3,A_3)\right).\end{array}$$

From our previous proposition (Liu et al., 2006), we know that $\alpha_{1,1}$, $\beta_{1,-1}$, $\beta_{-1,-1}$, the missing probabilities for marker B, and the haplotype frequencies are identifiable if and only if there is LD between the pseudo-biallelic marker A and the biallelic marker B. We also have the following expression:

$$\begin{aligned}\beta_{1,-1} &= \Pr(G=??|H=(A_1,A_{-1})) \\ &= \frac{\Pr(G=??,H=(A_1,A_2))+\Pr(G=??,H=(A_1,A_3))}{\Pr(H=(A_1,A_2))+\Pr(H=(A_1,A_3))} = \frac{p_{A_2}\alpha_{1,2}+p_{A_3}\alpha_{1,3}}{p_{A_2}+p_{A_3}}.\end{aligned}$$

Similarly, if we “pool” alleles {A$_1$, A$_3$} at marker A and denote it as A$_{-2}$, parameters $\alpha_{2,2}$, $\beta_{2,-2}$, $\beta_{-2,-2}$ and the haplotype frequencies are identifiable. In addition, we have

$$\beta_{2,-2}=\Pr(G=??|H=(A_2,A_{-2}))=\frac{p_{A_1}\alpha_{1,2}+p_{A_3}\alpha_{2,3}}{p_{A_1}+p_{A_3}}.$$

If we “pool” alleles {A$_1$, A$_2$} at marker A and denote it as A$_{-3}$, parameters $\alpha_{3,3}$, $\beta_{3,-3}$, $\beta_{-3,-3}$ and the haplotype frequencies are identifiable. In addition, we have

$$\beta_{3,-3}=\Pr(G=??|H=(A_3,A_{-3}))=\frac{p_{A_1}\alpha_{1,3}+p_{A_2}\alpha_{2,3}}{p_{A_1}+p_{A_2}}.$$

From the above procedure we can see that p_{A1}, p_{A2}, p_{A3}, $\beta_{1,-1}$, $\beta_{2,-2}$, and $\beta_{3,-3}$ can be identified uniquely (p_{A1}, p_{A2}, p_{A3} can be obtained from haplotype frequencies). Thus,we have the following equations:

$$\begin{bmatrix} \frac{p_{A_2}}{p_{A_2}+p_{A_3}} & \frac{p_{A_3}}{p_{A_2}+p_{A_3}} & 0 \\ \frac{p_{A_1}}{p_{A_1}+p_{A_3}} & 0 & \frac{p_{A_3}}{p_{A_1}+p_{A_3}} \\ 0 & \frac{p_{A_1}}{p_{A_1}+p_{A_2}} & \frac{p_{A_2}}{p_{A_1}+p_{A_2}} \end{bmatrix} \begin{bmatrix} \alpha_{1,2} \\ \alpha_{1,3} \\ \alpha_{2,3} \end{bmatrix} = \begin{bmatrix} \beta_{1,-1} \\ \beta_{2,-2} \\ \beta_{3,-3} \end{bmatrix}. \qquad \text{(P1)}$$

The determinant of the above equations is

$$\begin{vmatrix} \frac{p_{A_2}}{p_{A_2}+p_{A_3}} & \frac{p_{A_3}}{p_{A_2}+p_{A_3}} & 0 \\ \frac{p_{A_1}}{p_{A_1}+p_{A_3}} & 0 & \frac{p_{A_3}}{p_{A_1}+p_{A_3}} \\ 0 & \frac{p_{A_1}}{p_{A_1}+p_{A_2}} & \frac{p_{A_2}}{p_{A_1}+p_{A_2}} \end{vmatrix} = -\frac{2p_{A_1}p_{A_2}p_{A_3}}{\left(p_{A_1}+p_{A_2}\right)\left(p_{A_1}+p_{A_3}\right)\left(p_{A_2}+p_{A_3}\right)}.$$

It is not equal to zero if the allele frequencies are not zero, which is assumed, and therefore $\alpha_{1,2}$, $\alpha_{1,3}$, and $\alpha_{2,3}$ can be identified uniquely. This proves the proposition in the case of a tri-allelic marker and a biallelic marker.

Note that it is possible that when pooling alleles A$_r$ and A$_s$ the resulting pseudo-biallelic marker A may be in LE with marker B. We can show, however, that this may happen for at most one pair of alleles {Ar, As} at marker A. Assume that the pooling of alleles {A$_1$, A$_3$} and the pooling of alleles {A$_1$, A$_2$} both lead to LE with marker B. Then we should have

$$p_{A_2B_1}=p_{A_2}\cdot p_{B_1} \qquad (1)$$

$$p_{A_2B_2} = p_{A_2} \cdot p_{B_2} \tag{2}$$

$$p_{A_{-2}B_1} = p_{A_1B_1} + p_{A_3B_1} = p_{A_{-2}} \cdot p_{B_1} = \left(p_{A_1} + p_{A_3}\right) \cdot p_{B_1} \tag{3}$$

$$p_{A_{-2}B_2} = p_{A_1B_2} + p_{A_3B_2} = p_{A_{-2}} \cdot p_{B_2} = \left(p_{A_1} + p_{A_3}\right) \cdot p_{B_2} \tag{4}$$

$$p_{A_3B_1} = p_{A_3} \cdot p_{B_1} \tag{5}$$

$$p_{A_3B_2} = p_{A_3} \cdot p_{B_2} \tag{6}$$

From (3) and (5) above, we have $p_{A_1B_1} = p_{A_1} \cdot p_{B_1}$, and from (4) and (6) we have $p_{A_1B_1} = p_{A_1} \cdot p_{B_1}$. Together with (1), (2), (5), and (6), this means that markers A and B are in LE, which contradicts the condition of the proposition.

Without loss of generality, we assume that only pooling alleles $\{A_1, A_2\}$ leads the pseudo-biallelic marker A to be in LE with marker B. Thus, from pooling alleles $\{A_1, A_3\}$ and alleles $\{A_2, A_3\}$ we know that $\alpha_{1,1}$, $\beta_{1,-1}$, $\beta_{-1,-1}$, $\alpha_{2,2}$, $\beta_{2,-2}$, $\beta_{-2,-2}$ and that the haplotype frequencies of A_1B_1, A_1B_2, A_2B_1, A_2B_2 can be identified. Therefore, all the allele frequencies can be identified. From the observed data we know that

$$\Pr\left(O = (A_3A_{-3})\right) = \Pr\left(O = (A_3A_1)\right) + \Pr\left(O = (A_3A_2)\right) = 2p_{A_3}p_{A_{-3}}\left(1 - \beta_{3,-3}\right) = 2p_{A_3}\left(p_{A_1} + p_{A_2}\right)\left(1 - \beta_{3,-3}\right),$$

which means that $\beta_{3,-3}$ can be identified uniquely. Therefore, we still have equations (P1) and the consequent conclusion as shown above. Actually, knowing all of the allele frequencies allows us to obtain $\alpha_{r,s}$ at marker A with the observed data:

$$\Pr\left(O = (A_rA_s)\right) = \begin{cases} 2p_{A_r}p_{A_s}\left(1 - \alpha_{r,s}\right) & if \quad r \neq s \\ p_{A_r}p_{A_s}\left(1 - \alpha_{r,s}\right) & if \quad r = s. \end{cases}$$

Step 2: We have proved the proposition for M = 3 and N = 2. Now, assume that the proposition holds for M = K ≥ 3 and N = 2. Based on this, we now prove that the proposition holds for M = K + 1 and N = 2. For any missing probability related to genotype A_uA_v (denoted as $\alpha_{u,v}$), pool any two alleles other than A_u and A_v to "form" a pseudo-allele. Now, marker A has K alleles. From the above assumption, we know that the proposition holds in this case, so $\alpha_{u,v}$ can be identified. A similar procedure works for haplotype frequencies.

Much like the case of M = 3, pooling alleles $\{A_r, A_s\}$ may lead the pseudo-marker A to be in LE with marker B. Similarly, we can show that for $\{r,s : r \neq s$, and $r,s = 1, ---, M\}$ at most one pair of {r, s} may lead the pseudo-marker A to be in LE with marker B, after pooling alleles $\{A_r, A_s\}$. Therefore, all of the parameters except $\alpha_{r,s}$ can be identified. Observe that Pr(O =

$(A_r A_s)) = 2p_{A_r} p_{A_s} (1–\alpha_{r,s})$, where $\alpha_{r,s}$ can be obtained. So the proposition holds for any number M > 1 and N = 2.

Step 3: For the most general case, the proof is similar to the above for N = 2. Specifically, we can first prove the proposition for any number M and N = 3. Then we assume that the proposition holds for M and N = K ≥ 3. Similar to the above, we can show that the proposition holds for N = K + 1, which proves the proposition.□

2.3 Haplotype Frequency Estimation

The observed data likelihood is

$$\begin{aligned} L_{OBS} &= \prod_g [\Pr(G=g)]^{n_g} = \prod_g \left[\sum_{(h,h')\in S(g)} \Pr(G=g, H=(h,h')) \right]^{n_g} \\ &= \prod_g \left[\sum_{(h,h')\in S(g)} p_h p_{h'} \Pr(G=g|H=(h,h')) \right]^{n_g}, \end{aligned}$$

where the notation was defined before. When the two markers are in LD, the parameters are identifiable, and they can be estimated using the EM algorithm in a straightforward way (Liu et al., 2006).

3. Simulation Results

In our simulations, the first marker was simulated to have three alleles and the second to have two alleles. We have compared the performance of our method with HAPLO.STATS (Sinnwell, Schaid, Roland, & Yu, 2007) in estimating haplotype frequencies using simulated data. When there were no missing data, or when the data were missing at random, both methods gave accurate estimates of haplotype frequencies (Tables 3 and 4). When there were informatively missing data, our method could give accurate estimates of haplotype frequencies and missing data probabilities, whereas HAPLO.STATS gave biased estimates for haplotype frequencies. Table 5 shows the results from our method and HPALO.STATS from one simulation study in which 1,000 data sets were simulated, each with a sample size 1,000. Haplotype frequencies were set as 0.35, 0.15, 0.05, 0.1, 0.25, and 0.1. The missing probabilities were set as (0.22, 0, 0, 0, 0, 0) for the first marker and (0.16, 0, 0) for the second marker.

We noticed in our simulation study that when there are no missing data the performance of our method and that of HAPLO.STATS are almost identical (the estimates from the two methods only differ after the fifth decimal places). When there are missing data (missing at random or not), the point estimates are similar (when missing is at random), but the standard deviations from our method are usually larger than those from HAPLO.STATS. This is expected because our method includes more parameters than HAPLO.STATS does. We also noticed that the estimates of haplotype frequencies from our model are more accurate, with smaller variation than those of missing probabilities.

4. Application to Scleroderma Data

Systemic sclerosis (SSC), also called scleroderma, is a rare, chronic, autoimmune disease that afflicts an estimated 150,000 to 500,000 Americans, primarily females who are 30 to 50 years old at onset (Mayes, Varga, Buch, & Seibold, 2007). Clinically, the disease is divided into two major subtypes: diffuse cutaneous SSc (dcSSc) and limited cutaneous SSc (lcSSc). The diffuse cutaneous subtype is the most severe form and is generally more life threatening. It has a rapid onset, involves more widespread skin hardening, and is generally associated with significant internal organ involvement, especially renal crisis and diffuse alveolitis of the lung, along with anti-topoisomerase (anti-topo) auto-antibody. The limited cutaneous subtype (also called CREST Syndrome) is much milder and has a slow onset and progression. It is distinguished by Raynaud's phenomenon, telangiectasias, pulmonary hypertension, and the presence of anticentromere antibody (ACA). However, there is significant overlap in both the clinical manifestations and the specific auto-antibodies that occur in these two subtypes. It is not known what predisposes a susceptible individual to develop one subtype versus another, nor is there significant information about how the two disease subtypes may be pathogenically related. Studies have shown that genetic factors influence both the incidence and clinical manifestations of scleroderma (Arnett et al., 2001; Assassi & Tan, 2005; Tan et al., 1998; Wu et al., 2006). In a recent study, we analyzed a DNA repository for known, functional polymorphisms in the gene for the innate cytokine, macrophage, migration-inhibitory factor (MIF) (Baugh et al., 2002; Wu et al., 2006).

The data set contains genetic and clinical data collected from the Scleroderma Family Registry and DNA Repository at the University of Texas Health Science Center at Houston. A total of 740 subjects were studied; 203 of them had diffuse cutaneous SSc (dcSSc), 283 had limited cutaneous SSc (lcSSc), and the remaining 254 healthy subjects served as controls. Of the total sample, 655 of the subjects were white and 532 were female. A $\text{CATT}_{(5-8)}$ tetranucleotide repeat polymorphism at position −794 (with 5, 6, 7, and 8 repeats—i.e. four alleles) and a G/C single-nucleotide polymorphism (SNP) at position −173 (with two alleles) in the 5' promoter region of *MIF* had been genotyped for each subject.

There were 34 missing genotypes of CATT repeats at position −794, and 18 missing genotypes of SNP at position −173. For the CATT tetranucleotide repeat, there were 11 (4.33%) missing genotypes in controls, 16 (5.65%) in the lcSSc patient group, and 7 (3.45%) in the dcSSc patient group. For the SNP, there were 2 (0.79%) missing genotypes in controls, 13 (4.59%) in the lcSSc patient group, and 3 (1.48%) in the dcSSc patient group. There were no partial missing genotypes. Analysis did not find evidence of a haplotype effect (for the two closely linked polymorphisms described above) between cases and controls; however, there was evidence of a haplotype effect between the dcSSc patient group and the lcSSc patient group. More details about the data and the study have been published earlier (Wu et al., 2006).

The data were originally analyzed with HPlus 2.1.1, with missing at random being assumed for missing genotypes (Li, Khalid, Carlson, & Zhao, 2003; Wu et al., 2006). To be fair, to date there is no haplotype analysis method that can handle informatively missing genotypes for multi-allelic markers with samples of unrelated individuals, although there is no reason to believe that the missing genotypes are actually missing at random. In this study, we re-analyzed the data with HAPLO.STATS and with our proposed method. Because there was a very small number of 8 CATT repeats at the −794 CATT repeat polymorphism (4 in controls, 1 in patients with dcSSc, 0 in patients with lcSSc), we pooled 7 and 8 repeats together (which results in three alleles at this marker). We employed a commonly used likelihood test (Zaykin et al., 2002b; Zhao, Curtis, & Sham, 2000), and the significance level was calculated by re-sampling (Zaykin et al., 2002b; J. H. Zhao et al., 2000). Tables 6 and 7 provide the results for the case group vs. the control group and for the dcSSc patient group vs. the lcSSc patient group. Not surprisingly, the results from HAPLO.STATS confirmed the original findings. For cases vs. controls, the new method had the same results. However, for the dcSSc patient group vs. the lcSSc patient group, the new method gave different results. Missing probability estimates from our method are shown in Table 8. Obviously the missing pattern was far from random for this study sample, within and between different subject groups.

5. Summary and Final Notes

Missing genotypes are commonly encountered in genetic studies. The majority of statistical methods have usually been developed under the assumption of missing at random, which may be too simplistic and even invalid in many situations. There have been some studies investigating informative missingness (Allen, Rathouz, & Satten, 2003; Chen, 2004), but they considered different study designs involving related individuals. We have demonstrated that use of the missing at random assumption may lead to biased results. In this work, we have proposed to model informatively missing genotypes in haplotype analysis. We have analytically characterized the relationship between the identifiability of model parameters and LD between the markers under study, and we have shown that the presence of LD allows us to uniquely identify all of the parameters in this general missing data model, together with haplotype frequency estimates. We have also used Monte Carlo simulations to evaluate the proposed method. All simulation studies showed that our proposed methods perform well. We finish our considerations with a few remarks.

Remark 5.1. There are generally two ways to deal with missing data in practice: incorporating the missing data mechanism into the analysis (as is done in this work), or imputing the missing values. The advantage of the first approach is obvious. In many situations, however, there may be a need to impute the missing values. For example, the principal component analysis cannot deal with missing data.

Remark 5.2. Given the availability of our model and the fact that missing at random is a special case in our model, it is always safer to analyze the data under this general model for markers in LD, and our model also allows the missing-at-random hypothesis to be tested. However, we noticed in our study that the standard errors of the estimates from our method are usually larger than those from HAPLO.STATS (e.g., Table 4). This is expected because our method includes more parameters. As a consequence, our method may have lower power, although the power loss may be small (Liu et al., 2006). This is common in statistics and is the price that needs to be paid in order to use more complicated models. Prior to analysis, we suggest that the users explore their data first, such as looking at the missing rates and distributions. If the missing rates are low and/or the missing distributions do not seem to have obvious patterns, it may be acceptable to use simple models with the random-missing assumption. In addition, users can test whether the missingness is random or not using the new method and then choose among the models accordingly. Model selection criteria, such as AIC and BIC, can also be used to help in choosing the appropriate methods for the data.

Remark 5.3. Although our method failed to find evidence of a haplotype effect in either the scleroderma cases vs. controls or the dcSSc patients vs. the lcSSc patients, we do not conclude that there are no haplotype effects. There may be some confounders that mask the real signal, or the sample size may not be large enough for our method to detect the real effect. Further studies are needed to confirm these findings.

Remark 5.4. Although our proposed method is attractive, there are still limitations. We have dealt with only two markers in our study, although it is indispensable to analyze more markers in practice. Although the extension to more markers is computationally challenging, it is straightforward in principle. Some numerical techniques can be used to control the increased computation costs. For example, the partition-ligation method (Niu et al., 2002) is a potential way to reduce computing burden.

Remark 5.5. Although our missing data model is much more general than those currently in use, we do assume independent missingness at markers and HWE. These assumptions are reasonable in general. However, under certain circumstances they may be violated and may need to be relaxed for more general use of our method. In this paper, we have assumed that all genotyped data are correct without error. However, just like missing data, genotyping errors are commonly encountered in practice as well. Handling genotyping errors represents another promising future direction for research.

Acknowledgments

We thank Jason P. Sinnwell for his detailed explanation of HAPLO.STATS. This work was supported in part by NIH grants P30DK056336 (NL), and GM081488 (NL), AR049610 (RB), AR050498 (RB), GM59507 (HZ), GM57672 (HZ), from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the Associate Editor and two anonymous reviewers for their valuable comments and suggestions that improved the manuscript.

References

Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: What do we gain? Eur J Hum Genet 2001;9(4):291–300. [PubMed: 11313774]

Allen AS, Rathouz PJ, Satten GA. Informative missingness in genetic association studies: Case-parent designs. Am J Hum Genet 2003;72(3):671–680. [PubMed: 12592606]

Arnett FC, Cho M, Chatterjee S, Aguilar MB, Reveille JD, Mayes MD. Familial occurrence frequencies and relative risks for systemic sclerosis (scleroderma) in three United States cohorts. Arthritis Rheum 2001;44(6):1359–1362. [PubMed: 11407695]

Assassi S, Tan FK. Genetics of scleroderma: Update on single nucleotide polymorphism analysis and microarrays. Curr Opin Rheumatol 2005;17(6):761–767. [PubMed: 16224255]

Baugh JA, Chitnis S, Donnelly SC, Monteiro J, Lin X, Plant BJ, Wolfe F, Gregersen PK, Bucala R. A functional promoter polymorphism in the macrophage migration inhibitory factor (MIF) gene associated with disease severity in rheumatoid arthritis. Genes Immun 2002;3(3):170–176. [PubMed: 12070782]

Botstein D, Risch N. Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease. Nat Genet 2003;33(Suppl):228–237. [PubMed: 12610532]

Chen YH. New approach to association testing in case-parent designs under informative parental missingness. Genet Epidemiol 2004;27(2):131–140. [PubMed: 15305329]

Chiano MN, Clayton DG. Fine genetic mapping using haplotype analysis and the missing data problem. Ann Hum Genet 1998;62(Pt 1):55–60. [PubMed: 9659978]

Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 1990;7(2):111–122. [PubMed: 2108305]

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen MM, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S. Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. Bioinformatics 2005;21 (9):1958–1963. [PubMed: 15657097]

Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 2003;73(6):1316–1329. [PubMed: 14631556]

Excoffier L, Laval G, Balding D. Gametic phase estimation over large genomic regions using an adaptive window approach. Hum Genomics 2003;1(1):7–19. [PubMed: 15601529]

Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12(5):921–927. [PubMed: 7476138]

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE locus variation and Alzheimer's disease. Genome Res 2001;11(1):143–151. [PubMed: 11156623]

Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 2000;67(4):947–959. [PubMed: 10954684]

Hao K, Cawley S. Differential dropout among SNP genotypes and impacts on association tests. Hum Hered 2007;63(3–4):219–228. [PubMed: 17347569]

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 1997;60(4): 772–789. [PubMed: 9106523]

Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. Springer; New York: 2001.

Hawley ME, Kidd KK. HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 1995;86(5):409–411. [PubMed: 7560877]

Kalinowski ST, Hedrick PW. Estimation of linkage disequilibrium for loci with multiple alleles: Basic approach and an application using data from bighorn sheep. Heredity 2001;87(Pt 6):698–708. [PubMed: 11903565]

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. Am J Hum Genet 2000;66(6): 1882–1899. [PubMed: 10788337]

Li SS, Khalid N, Carlson C, Zhao LP. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. Biostatistics 2003;4(4):513–522. [PubMed: 14557108]

Lin DY, Zeng D, Millikan R. Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet Epidemiol 2005;29(4):299–312. [PubMed: 16240443]

Liu N, Beerman I, Lifton R, Zhao H. Haplotype analysis in the presence of informatively missing genotype data. Genet Epidemiol 2006;30(4):290–300. [PubMed: 16528706]

Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 1995;56(3):799–810. [PubMed: 7887436]

Mayes, MD.; Varga, J.; Buch, MH.; Seibold, JR. Systemic Sclerosis.. In: Klippel, JH.; Stone, JH.; Crofford, LJ.; White, PH., editors. Primer on the Rheumatic Diseases. Vol. 13th ed.. Springer; 2007. p. 343-362.

Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 2002;23(3):221–233. [PubMed: 12384975]

Nicolae DL, Wu X, Miyake K, Cox NJ. GEL: A novel genotype calling algorithm using empirical likelihood. Bioinformatics 2006;22(16):1942–1947. [PubMed: 16809396]

Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 2002;70(1):157–169. [PubMed: 11741196]

Nothnagel M, Furst R, Rohde K. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered 2002;54(4):186–198. [PubMed: 12771551]

Schaid DJ. Genetic epidemiology and haplotypes. Genet Epidemiol 2004a;27(4):317–320. [PubMed: 15543637]

Schaid DJ. Evaluating associations of haplotypes with traits. Genet Epidemiol 2004b;27(4):348–364. [PubMed: 15543638]

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70(2):425–434. [PubMed: 11791212]

Sinnwell, J.; Schaid, D.; Roland, C.; Yu, Z. haplo.stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.3.1. 2007. Retrieved, from the World Wide Web: http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm

Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 2005;76(3):449–462. [PubMed: 15700229]

Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68(4):978–989. [PubMed: 11254454]

Tan FK, Stivers DN, Foster MW, Chakraborty R, Howard RF, Milewicz DM, Arnett FC. Association of microsatellite markers near the fibrillin 1 gene on human chromosome 15q with scleroderma in a Native American population. Arthritis Rheum 1998;41(10):1729–1737. [PubMed: 9778214]

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661–678. [PubMed: 17554300]

Tishkoff SA, Pakstis AJ, Stoneking M, Kidd JR, Destro-Bisol G, Sanjantila A, Lu RB, Deinard AS, Sirugo G, Jenkins T, Kidd KK, Clark AG. Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: Implications for modern human origins. Am J Hum Genet 2000;67(4):901–925. [PubMed: 10986042]

Tregouet DA, Escolano S, Tiret L, Mallet A, Golmard JL. A new algorithm for haplotype-based association analysis: The Stochastic-EM algorithm. Ann Hum Genet 2004;68(Pt 2):165–177. [PubMed: 15008795]

Wijsman EM. A deductive method of haplotype analysis in pedigrees. Am J Hum Genet 1987;41(3):356–373. [PubMed: 3115093]

Wu SP, Leng L, Feng Z, Liu N, Zhao H, McDonald C, Lee A, Arnett FC, Gregersen PK, Mayes MD, Bucala R. Macrophage migration inhibitory factor promoter polymorphisms and the clinical expression of scleroderma. Arthritis Rheum 2006;54(11):3661–3669. [PubMed: 17075815]

Yu Z, Schaid DJ. Methods to impute missing genotypes for population data. Hum Genet 2007a;122(5): 495–504. [PubMed: 17851696]

Yu Z, Schaid DJ. Sequential haplotype scan methods for association analysis. Genet Epidemiol 2007b; 31(6):553–564. [PubMed: 17487883]

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 2002a;53(2):79–91. [PubMed: 12037407]

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 2002b;53(2):79–91. [PubMed: 12037407]

Zhang K, Sun F, Zhao H. HAPLORE: A program for haplotype reconstruction in general pedigrees without recombination. Bioinformatics 2005;21(1):90–103. [PubMed: 15231536]

Zhang P, Sheng H, Morabia A, Gilliam TC. Optimal step length EM algorithm (OSLEM) for the estimation of haplotype frequency and its application in lipoprotein lipase genotyping. BMC Bioinformatics 2003;4(1):3. [PubMed: 12529185]

Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK. Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 2000;67(4):936–946. [PubMed: 10968775]

Zhao JH, Curtis D, Sham PC. Model-free analysis and permutation tests for allelic associations. Hum Hered 2000;50:133–139. [PubMed: 10799972]

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M. An arabidopsis example of association mapping in structured samples. PLoS Genet 2007;3(1):e4. [PubMed: 17238287]

Zhao LP, Li SS, Khalid N. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 2003;72(5):1231–1250. [PubMed: 12704570]

Table 1

Genotype Penetrances for a Tri-Allelic Marker

	A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_2A_3	A_3A_3	??
A_1A_1	1- $\alpha_{1,1}$	0	0	0	0	0	$\alpha_{1,1}$
A_1A_2	0	1- $\alpha_{1,2}$	0	0	0	0	$\alpha_{1,2}$
A_1A_3	0	0	1- $\alpha_{1,3}$	0	0	0	$\alpha_{1,3}$
A_2A_2	0	0	0	1- $\alpha_{2,2}$	0	0	$\alpha_{2,2}$
A_2A_3	0	0	0	0	1- $\alpha_{2,3}$	0	$\alpha_{2,3}$
A_3A_3	0	0	0	0	0	1- $\alpha_{3,3}$	$\alpha_{3,3}$

NOTE: The column headings are the possible observed genotypes. The row headings are the true genotypes. “?” denotes a missing allele. Each entry is the probability of observing the observed genotype given the true genotype.

Table 2

Genotype Penetrances for a Tri-Allelic Marker after "Pooling" Two Alleles to Form a Pseudo-Biallelic Marker

	A_1A_1	**A_1A_{-1}**	**$A_{-1}A_{-1}$**	**??**
A_1A_1	$1- \alpha_{1,1}$	0	0	$\alpha_{1,1}$
A_1A_{-1}	0	$1- \beta_{1,-1}$	0	$\beta_{1,-1}$
$A_{-1}A_{-1}$	0	0	$1- \beta_{-1,-1}$	$\beta_{-1,-1}$

NOTE: The column headings are the possible observed genotypes. The row headings are the true genotypes. Each entry is the probability of observing the observed genotype given the true genotype.

Table 3

Results from Simulations without Missing Genotypes

	p_{h_1}	p_{h_2}	p_{h_3}	p_{h_4}	p_{h_5}	p_{h_6}
True	0.35	0.15	0.05	0.1	0.25	0.1
			HAPLO.STATS			
Mean	0.3502	0.1500	0.0499	0.1002	0.2495	0.1001
Median	0.3507	0.1502	0.0500	0.0999	0.2495	0.0997
Standard deviation	0.0118	0.0091	0.0057	0.0072	0.0108	0.0081
			New Method			
Mean	0.3502	0.1500	0.0499	0.1002	0.2495	0.1001
Median	0.3507	0.1502	0.0500	0.0999	0.2495	0.0997
Standard deviation	0.0118	0.0091	0.0057	0.0072	0.0108	0.0081

NOTE: There were no missing genotypes in the data. ph_1 , ph_2 ,... , ph_6 are the frequencies for haplotypes A_1B_1, A_1B_2, A_2B_1, A_2B_2, A_3B_1, A_3B_2, respectively. The results were based on 1,000 simulations with a sample size of 1,000 for each simulated data set. For the new method, the estimates of missing probabilities are all zero.

Table 4

Results from Simulations with Randomly Missing Genotypes

	p_{h_1}	p_{h_2}	p_{h_3}	p_{h_4}	p_{h_5}	p_{h_6}
True	0.35	0.15	0.05	0.1	0.25	0.1
			HAPLO.STATS			
Mean	0.3510	0.1497	0.0501	0.0998	0.2493	0.1001
Median	0.3511	0.1494	0.0499	0.1000	0.2495	0.1002
Standard deviation	0.0124	0.0099	0.0061	0.0077	0.0111	0.0088
			New Method			
Mean	0.3479	0.1491	0.0504	0.1010	0.2506	0.1010
Median	0.3470	0.1491	0.0503	0.1009	0.2503	0.1006
Standard deviation	0.0174	0.0119	0.0072	0.0095	0.0162	0.0100

NOTE: The missing probabilities were all set to 0.1. ph_1 , ph_2 ,... , ph_6 are the frequencies of haplotypes A_1B_1, A_1B_2, A_2B_1, A_2B_2, A_3B_1, A_3B_2, respectively. The results were based on 1,000 simulations with a sample size of 1,000 for each simulated data set. For the new method, the estimates of missing probabilities are (0.0876, 0.0939, 0.0963, 0.1374, 0.1117, 0.1103) and (0.0959, 0.0992, 0.1081) for the first and second markers, with standard deviations (0.0588, 0.0664, 0.0392, 0.1378, 0.0787, 0.0824) and (0.0408, 0.0322, 0.0772), respectively.

Table 5

Results from Simulations with Informatively Missing Genotypes

	p_{h_1}	p_{h_2}	p_{h_3}	p_{h_4}	p_{h_5}	p_{h_6}
True	0.35	0.15	0.05	0.1	0.25	0.1
			HAPLO.STATS			
Mean	0.3204	0.1509	0.0497	0.1090	0.2568	0.1131
Median	0.3201	0.1513	0.0495	0.1091	0.2571	0.1135
Standard deviation	0.0121	0.0099	0.0058	0.0080	0.0116	0.0094
			New Method			
Mean	0.3436	0.1468	0.0504	0.1025	0.2543	0.1024
Median	0.3445	0.1468	0.0502	0.1026	0.2534	0.1027
Standard deviation	0.0153	0.0106	0.0058	0.0080	0.0146	0.0092

NOTE: The missing probabilities were set as (0.22, 0, 0, 0, 0, 0) for the first marker, and (0.16, 0, 0) for the second marker. p_{h_1} , p_{h_2} ,... , p_{h_6} are the frequencies of haplotypes A_1B_1, A_1B_2, A_2B_1, A_2B_2, A_3B_1, A_3B_2, respectively. The results were based on 1,000 simulations with a sample size of 1,000 for each simulated data set. For the new method, the estimates of missing probabilities are (0.1793, 0.0084, 0.0066, 0.0336, 0.0174, 0.0366) and (0.1537, 0.0045, 0.0081) for the first and second markers, with standard deviations (0.0521, 0.0229, 0.0171, 0.0792, 0.0422, 0.0739) and (0.0122, 0.0117, 0.0222), respectively.

Table 6

Comparison of Results from the Scleroderma Data: Cases vs. Controls

	p_{h_1}	p_{h_2}	p_{h_3}	p_{h_4}	p_{h_5}	p_{h_6}
HAPLO.STATS						
Cases	0.2566	0.0157	0.5623	0.0473	0.0046	0.1134
Controls	0.2353	0.0179	0.5690	0.0444	0.0022	0.1312
Likelihood ratio statistic		2.0078		P value	0.880	
New Method						
Cases	0.2652	0.0160	0.5533	0.0447	0.0000	0.1209
Controls	0.2473	0.0185	0.5559	0.0429	0.0000	0.1355
Likelihood ratio statistic		1.2301		P value	0.964	

NOTE: The results were from the haplotype analysis of 486 scleroderma cases and 254 healthy controls. The column heading (ph_i 's) denotes the frequencies of the haplotypes that appeared in the data. The entries with row headings “cases” and “controls” are estimates of haplotype frequencies for cases and controls, respectively. The entries with “likelihood ratio statistic” are likelihood ratio test statistics and associated P values. The P values were calculated based on 1,000 simulations.

Table 7

Comparison of Results from the Scleroderma Data: lcSSc vs. dcSSc

	p_{h_1}	p_{h_2}	p_{h_3}	p_{h_4}	p_{h_5}	p_{h_6}
			HAPLO.STATS			
lcSSc patient group	0.2781	0.0014	0.5861	0.0435	0.0041	0.0868
dcSSc patient group	0.2365	0.0254	0.5201	0.0623	0.0055	0.1502
Likelihood ratio statistic		19.4628		P value	0.006	
			New Method			
lcSSc patient group	0.2893	0.0004	0.5735	0.0420	0.0000	0.0948
dcSSc patient group	0.2424	0.0254	0.5137	0.0588	0.0000	0.1597
Likelihood ratio statistic		19.5641		P value	0.258	

NOTE: The results were from the haplotype analysis of 283 limited cutaneous SSc (lcSSc) patients and 203 diffuse cutaneous SSc (dcSSc) patients. The column heading (p_{h_i}'s) denotes the frequencies of the haplotypes that appeared in the data. The entries with row headings “lcSSc patient group” and “dcSSc patient group” are estimates of haplotype frequencies for the lcSSc patients and dcSSc patients, respectively. The entries with “likelihood ratio statistic” are likelihood ratio test statistics and associated P values. The P values were calculated based on 1,000 simulations.

Table 8

Estimates of Missing Probabilities from the Scleroderma Data

	CATT Tetranucleotide Repeat						SNP		
	$\alpha_{1,1}$	$\alpha_{1,2}$	$\alpha_{1,3}$	$\alpha_{2,2}$	$\alpha_{2,3}$	$\alpha_{3,3}$	$\gamma_{1,1}$	$\gamma_{1,2}$	$\gamma_{2,2}$
Controls	0.271	0.042	0.000	0.000	0.068	0.000	0.012	0.000	0.000
Cases	0.000	0.085	0.258	0.000	0.025	0.014	0.035	0.034	0.000
lcSSc patient group	0.112	0.067	0.124	0.000	0.147	0.000	0.040	0.067	0.000
dcSSc patient group	0.000	0.062	0.291	0.000	0.000	0.038	0.025	0.000	0.000

NOTE: The entries are estimates of missing probabilities in the case group, the control group, the limited cutaneous SSc (lcSSc) patient group, and the diffuse cutaneous SSc (dcSSc) patient group at the two markers.