Search for Haplotype-Interactions That Influence Susceptibility to Type 1 Diabetes Using Unphased Genotype Data

Jian Zhang[1]

Institute of Mathematics and Statistics, University of Kent at Canterbury

Kent CT2 7NF, UK; EURANDOM, Eindhoven, The Netherlands; and The Chinese

Academy of Sciences, Beijing, China

Faming Liang

Department of Statistics, Texas A& M University College Station, USA

Willem R.M. Dassen

Department of Cardiology, Maastricht University, Maastricht, The Netherlands

Pieter A. Doevendans

Interuniversitairy Cardiology Institute of The Netherlands, Utrecht, The Netherlands

and

Mathisca de Gunst

Department of Mathematics, Free University, Amsterdam, The Netherlands;

EURANDOM, Eindhoven, The Netherlands

Received ______________________; accepted ______________________

Running title: Haplotype interaction

[1] Address for correspondence: Dr. Jian Zhang, Institute of Mathematics and Statistics, University of Kent at Canterbury, Kent CT2 7NF, United Kingdom. Tel: +441227 823661; Fax: +441227 827932; E-mail: j.zhang@kent.ac.uk

ABSTRACT

Type 1 diabetes is a T-cell mediated chronic disease, characterised by the autoimmune destruction of pancreatic insulin-producing β-cells and complete insulin deficiency. It is the result of a complex interrelation of genetic and environmental factors, most of which have yet to be identified. Simultaneous identification of these genetic factors using unphased genotype data has received increasing attention in the past few years. Several approaches have been described, such as the modified transmission/disequilibrium test procedure, the conditional extended transmission disequilibrium test, and the stepwise logistic-regression procedure. These approaches are limited either by being restricted to family data or by ignoring so-called haplotype interactions between alleles. To overcome this limit, this report provides a general method to identify the haplotype blocks that interact to define the risk for a complex disease based on unphased genotype data. The principle underpinning the proposal is minimal entropy. The performance of our procedure is illustrated for both simulated and real data. In particular, for a set of Dutch type 1 diabetes data, our procedure suggests some novel evidence of the interactions between and within haplotype-blocks, which are across chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 15, 16, 17, 19, and 21. The results demonstrate that by considering interactions between potential disease haplotype blocks, we may succeed in identifying disease-predisposing genetic variants that might otherwise have remained undetected.

Introduction

Insulin-dependent diabetes mellitus (IDDM [MIM 222100]), or type 1 diabetes, is a common chronic disease characterized by autoimmune destruction of pancreatic β-cells and complete insulin deficiency (Cordell and Todd 1995; Friday et al. 1999; Schranz and Lernmark 1998). The importance of some genetic factors for the etiology of type 1 diabetes, such as HLA, has been established unequivocally, although their precise mechanism has not been identified. Evidence for a role of the immune system and apoptosis is accumulating. Both processes contribute to the deterioration of beta cells in the islets of Langerhans in the pancreas. Despite this information, in most patients no definite genetic cause can be determined, not even in the presence of a positive family history. In this paper we present a method for testing the influence of haplotype interactions on getting a disease when unphased genotypes are available for a number of cases and controls, and we apply this method to genotype data of type 1 diabetes patients and healthy controls. Here, as in Bugawan et al.(2003), the haplotype interaction is defined as the statistical dependence between alleles at different loci.

The increasing availability of polymorphic markers such as single nucleotide polymorphisms (SNPs), automated genotyping technology, and large collections of family (or case-control) based data have enabled the design of genome-wide screens for several populations. Such screens have led to the location of susceptibility loci for type 1 diabetes in various chromosomal regions, suggesting that type 1 diabetes is a multigenic disorder in the sense that onset of the disease requires the simultaneous presence of a subset of susceptibility genes. Most recent research efforts have been put on HLA genes (see Pugliese (2001) and Cox et al. (2001) for a review). The importance of the HLA class II haplotypes was shown by Noble et al.(2002) in families with at least two children with insulin dependent diabetes.

Once a disease-predisposing region has been localized, a number of potentially causative genetic variants may exist in the regions, including a large number of SNPs. Whereas for monogenetic diseases very often one base change in the coding region of a gene is sufficient to cause the disease, for multigenic diseases the effect of any single genetic variant on the risk of the disease may be small, which makes identification of these variants difficult (Drysdale et al. 2000). Furthermore, the following questions related to identification of the multiple risk variants arise. First, it is not clear which combination of variants has a causative role in the disease. Secondly, it remains unknown whether susceptibility for the disease arises due to the effects of these variants acting independently or due to some important interactions between the variants.

These questions have received increasing attention recently (see, for example, Bugawan et al. 2003; Cordell and Clayton 2002; Cox et al. 1999; Dassen et al. 2001; Valdes and Thomson 1997). Cordell and Clayton (2002) proposed a simple but powerful stepwise logistic-regression procedure, which allows for testing the dominance effects of different combinations of polymorphisms as well as genotype interactions in the analysis of case-control data. In particular, they measured genotype interactions in terms of penetrance for developing disease. However, haplotype interactions, in the sense that the underlying haplotype pairs of unphased genotypes may have different disease risks so that there are disease-predisposing interactions, cannot be dealt with in their approach. To see this, for the moment we consider two diallelic variants of interest in a region, variant 1 with one of the unphased genotypes aa, AA and aA, and variant 2 with one of the unphased genotypes bb, BB and bB. There are nine possible combinations (also called genotypes) observed at the two variants: $aa/bb, aa/bB, aa/BB, AA/bb, AA/bB, AA/BB$, $aA/bb, aA/bB$, and aA/BB, where, for example, aA/bb means that the alleles in variants 1 and 2 are $\{a, A\}$ and $\{b, b\}$, respectively. All these genotypes except for aA/bB can be uniquely decomposed into a pair of haplotypes. For aA/bB, there are two compatible possible haplotype pairs, $(a, b)/(A, B)$

and $(a, B)/(A, b)$. The pairing described here indicates that allele a is in coupling with allele b or allele a is coupling with allele B. It is only when these two haplotype pairs have different disease risks that there may be potential disease-predisposing interactions between a and b or a and B. As pointed out by a reviewer, even when the haplotype pairs do have different disease risks, it does not necessarily mean that the alleles interact in anything other than a statistical sense, since this phenomenon could occur if alleles a and b, say, were in linkage disequilibrium with (and thus marking a haplotype containing) another predisposing variant not included in the analysis. Note that the stepwise logistic-regression procedure takes genotypes as explanatory variables and so the possible difference between the effects of the underlying haplotypes on the disease is ignored.

An alternative test is called the haplotype method (Valdes and Thomson 1997), which compares the relative frequencies of alleles at a secondary locus on haplotypes that are identical at a primary locus (or loci). The problem with the haplotype method is that, often, the haplotypes are not known. Although one can statistically infer the haplotypes from unphased genotypes, it is unclear how to judge the significance of the results from the haplotype method if we want to take into account the possible haplotyping errors. Several other approaches have been described for simultaneous identification of genetic factors using unphased genotype data, such as the modified transmission/disequilibrium test procedure (Cucca et al. 2001) and the conditional extended transmission disequilibrium test (Koeleman et al. 2000). These approaches are also limited, either by being restricted to family data or to haplotype data. This and the fact that there are 2^{m-1} possible haplotype pairs for a genotype of m heterozygous sites, which results in a considerable number of potential haplotype interactions when m is large, motivated us to develop a special procedure for testing such interactions. The proposed method is based on minimal entropy, reflecting the principle that a good prediction of haplotype interactions should extract a maximum amount of information from data and thus most parsimoniously explain the underlying haplotype structure given unphased genotypes. In general, the computation of the entropy statistic is very intensive. To solve this problem, we have developed a new Markov chain Monte Carlo algorithm, called structure-annealing algorithm.

Two types of approaches for the investigation of interaction can be distinguished: those that consider interaction in the sense of linkage disequilibrium between closely linked loci (Wall and Pritchard 2003) and those that consider interaction in the sense of effects on disease risk (Cordell and Todd 1995; Cordell et al. 2001). In this paper we focus on the linkage disequilibrium approach while investigating interaction between all loci, and hence also between possibly unlinked loci. For any two haplotype-blocks, let us denote by p_{1a} and p_{2b} the probabilities of occurrence, respectively, for allele a at block 1 and for allele b at block 2. Let p_{ab} be the probability of simultaneous occurrence of a and b. We are trying to test whether for all a and b, $p_{ab} = p_{1a}p_{2b}$. We assess the evidence for interactions between and within (possibly unlinked) haplotype blocks on different chromosomal regions by using a permutation procedure. Since the strength of linkage disequilibrium pattern is not, typically, a monotonic function of recombination distance when there exist selective forces that favour certain haplotypes over others as might be the case for type 1 diabetes (Fain et al. 2001), we needed to develop an approach that is independent of this distance. Naturally, we are mainly interested in identification of disease-predisposing interactions by the comparisons between cases and controls. The disease-predisposing interactions are found in a second stage by contrasting the interaction patterns observed for patients with the interaction patterns observed for healthy controls. These interactions could facilitate understanding of the pathological mechanisms involved in the disease, as well as the further identification of some haplotype blocks that only provide significant association with the disease when their interactions with other blocks are taken into account.

As an illustration of our method, we present in this paper a re-analysis of a set of genotypes that was obtained from a cohort of 89 Dutch type 1 diabetes patients and 47 healthy controls, with a 65 polymorphisms detection assay originally designed for unraveling the multigenic cause of atherosclerosis (Dassen et al. 2001). Since both diabetes mellitus and atherosclerosis can be regarded as metabolic diseases, with many overlapping biochemical and clinical parameters, the variants that are susceptible to atherosclerosis may also be the cause of type 1 diabetes. Dassen and co-workers examined whether certain types of combinations of SNPs confer susceptibility to type 1 diabetes in the cohort by logistic regression and self-learning neural networks. They found that a set of four polymorphisms, could predict 79.9% of the cases correctly. However, a significant number of polymorphisms could not be interpreted by their method. Note that all these variants were selected from the pathways of lipid and homocysteine metabolism, regulation of blood pressure and coagulation, inflammation, cellular adhesion, and matrix integrity. So we wonder whether the variants that were unexplained in the above-mentioned study may serve as transitive (or supporting) variants in the sense that they interact with some etiological variants within and between these pathways.

Before we applied the proposed procedure to the above-mentioned Dutch type 1 diabetes data, we evaluated the power of our approach by conducting a simulation study in which four different combinations of mutation and recombination rates were considered. The results are presented below. They suggest that a high accuracy can be achieved if appropriate critical values for our entropy statistics are selected. Note that, although the coalescent model that we have used for our simulations, has been shown to be very helpful in modelling haplotype populations (Stephens et al. 2001), it is still not easy to statistically test whether this model fits to real data like the Dutch type 1 diabetes data. Therefore, the thresholds that were obtained from the simulations were used as a guide to the corresponding parameters as we applied our method to the data. The results of our data analysis show some evidence for a haplotype interaction network that is potentially associated with type 1 diabetes, and which includes the up-interactions between the haplotype blocks from the chromosomes pairs (1,4), (1,12), (1,19), (6, 7), and (17, 21); and the down-interactions between blocks from the chromosome pairs (2,7), (3,19), (5,7), (6,21), and (7,11). There are several other less significant pairs. Here, up-interaction (down-interaction) means there exists a significant increase (decrease) of interaction between two blocks for patients over that for controls. We further found some disease predisposing intra-block interactions on chromosomes 1, 6, 7, 8, and 11. Finally, we searched for loci-interactions that may account for these block-interactions. As a result, totally 25 potential disease-predisposing interactions between loci are predicted, which indicates 19 gene-gene interactions among 19 candidate genes. Having found four dominant variants (Dassen et al. 2001), we predicted from the interaction-network 19 transitive variants. Our results clearly demonstrate that by considering interactions between haplotype blocks, we may succeed in identifying disease-predisposing genetic variants that might otherwise have remained undetected.

Methods

Haplotype likelihood

Let $\boldsymbol{G} = (G_1, \ldots, G_n)^T$ denote the observed genotypes for n individuals from a population, where $G_i = (g_{i1}, \ldots, g_{iL})^T$, with g_{ij} is the genotype of individual i at locus j, and L the total number of observed loci per individual. For simplicity, let g_{ij} take values of 0, 1, or 2 for the cases where its genetic haplotype at the locus j is homozygous and identical with a pre-specified reference, homozygous but different from the reference, or heterozygous, respectively. In addition, we let $g_{ij} = 7$ if allele 0 is missing at locus j, $g_{ij} = 8$

if allele 1 is missing, and $g_{ij} = 9$ if both alleles are missing. A genotype is called ambiguous if it has at least two heterozygous sites. Let $\boldsymbol{H} = (H_1, \ldots, H_n)^T$, where $H_i = (H_{i1}, H_{i2})$ denotes the unobserved haplotype pair of G_i, $H_i \in \mathcal{H}_i$, the set of all possible haplotype pairs compatible to G_i. Given $\boldsymbol{G}$, under the assumption of Hardy-Weinberg equilibrium (Weir, Chapter 3, 1996), the 'haplotype-likelihood' can then be written as

$$L(\boldsymbol{G}|\mathbf{p}, \boldsymbol{H}) = \prod_{i=1}^{n} p(H_{i1})p(H_{i2}), \tag{1}$$

where $p(\cdot)$ denotes the population frequencies of the corresponding haplotype, and $\mathbf{p} = (p_1, \ldots, p_{m_0})$. Here we assume that totally there are m_0 possible haplotypes compatible to $\boldsymbol{G}$.

Haplotype entropy

While performing a haplotype inference, we are usually only interested in $\boldsymbol{H}$, and, hence, $\mathbf{p}$ works as a nuisance parameter in (1). Here we follow Zhang et al (2001) to eliminate the nuisance parameter by a maximization procedure, that is, we substitute $\mathbf{p}$ in (1) by its MLE (maximum likelihood estimate). Thus, we have the following profile log-likelihood,

$$l(\boldsymbol{G}|\boldsymbol{H}) = \sum_{k=1}^{k_0} \frac{s_k}{2n} \log \frac{s_k}{2n}.$$

where k_0 denotes the number of different haplotypes in $\boldsymbol{H}$, and $s_1, \ldots, s_{k_0}$ denote their respective frequencies. We define

$$S(\boldsymbol{H}) = -l(\boldsymbol{G}|\boldsymbol{H}),$$

where $S(\boldsymbol{H})$ is the entropy of the frequencies of different haplotypes in $\boldsymbol{H}$, and

$$s(\boldsymbol{G}) = \min\{S(\boldsymbol{H}) : \quad \boldsymbol{H} \text{ is compatible with } \boldsymbol{G}\}.$$

Note that $S(\boldsymbol{H})$ attains its minimum at $\hat{\boldsymbol{H}}$, the MLE of H in (1), so that

$$s(\boldsymbol{G}) = S(\hat{\boldsymbol{H}}). \tag{2}$$

For example, suppose that $\boldsymbol{G} = \{(0,0,0),(1,0,0),(2,2,0),(1,1,2)\}$. Then there are two possible ways to decompose these genotypes into haplotypes, namely $\boldsymbol{H}_1 = \{(0,0,0)/(0,0,0),(1,0,0)/(1,0,0),(1,0,0)/(0,1,0),(1,1,0)/(1,1,1)\}$ and $\boldsymbol{H}_2 = \{(0,0,0)/(0,0,0),(1,0,0)/(1,0,0),(1,1,0)/(0,0,0),(1,1,0)/(1,1,1)\}$. The corresponding values of the haplotype-likelihood shown in (1) are, respectively,

$$p((0,0,0))^2 p((1,0,0))^2 p((1,0,0))p((0,1,0))p((1,1,0))p((1,1,1)) \tag{3}$$

and

$$p((0,0,0))^2 p((1,0,0))^2 p((1,1,0))p((0,0,0))p((1,1,0))p((1,1,1)), \tag{4}$$

where the unknown population frequencies of the five different haplotypes in (3) satisfy the equation

$$p((0,0,0)) + p((1,0,0)) + p((0,1,0)) + p((1,1,0)) + p((1,1,1)) = 1. \tag{5}$$

and the unknown population frequencies of the four different haplotypes in (4) are constrained by

$$p((0,0,0)) + p((1,0,0)) + p((1,1,0)) + p((1,1,1)) = 1. \tag{6}$$

Given $\boldsymbol{H}_1$ and under the constraint (5), the maximum of the logarithm of the likelihood in (3) is given by $2/8\log(2/8) + 3/8\log(3/8) + 3/8\log(1/8) = -S(\boldsymbol{H}_1)$. Analogously, given $\boldsymbol{H}_2$ and under the constraint (6), the maximum of the logarithm of the likelihood in (3) is equal to $3/8\log(3/8) + 2/8\log(2/8) + 2/8\log(2/8) + 1/8\log(1/8) = -S(\boldsymbol{H}_2)$. Obviously, $S(\boldsymbol{H}_2) < S(\boldsymbol{H}_1)$. Hence $s(\boldsymbol{G}) = S(\boldsymbol{H}_2)$.

In this paper, we call $s(\boldsymbol{G})$ the haplotype entropy of $\boldsymbol{G}$. The quantity $s(\boldsymbol{G})$ measures the diversity of the underlying haplotypes compatible with $\boldsymbol{G}$ since the entropy is a well-known measure of variation for a system in information theory (Jones 1979). The stronger the interactions among the loci of $\boldsymbol{G}$, the less diverse the underlying haplotypes, and the smaller the value of $s(\boldsymbol{G})$. To explain this claim intuitively, we consider only three diallelic loci at which there are eight possible haplotypes, namely, $h_1 = (0,0,0)^T$, $h_2 = (0,0,1)^T$, $h_3 = (0,1,0)^T$, $h_4 = (0,1,1)^T$, $h_5 = (1,0,0)^T$, $h_6 = (1,0,1)^T$, $h_7 = (1,1,0)^T$, and $h_8 = (1,1,1)^T$. Let $p(h_i)$ be the population frequency of h_i for $1 \leq i \leq 8$. The population haplotype entropy, defined as $-\sum_{i=1}^{8} p(h_i)\log(p(h_i))$, is a measure of the diversity of the above haplotype population. In practice, we might only have a sample of genotypes of size n, say $\boldsymbol{G}$, which are assumed to be generated from these haplotypes according to the Hardy-Weinberg equilibrium. Then the haplotype entropy $S(\hat{\boldsymbol{H}})$ in (2) gives rise to an empirical version of the above population entropy. To see how the haplotype entropy changes as the strength of interaction (i.e., dependence) increases, we first calculate this entropy when there are no interactions among the three loci. In this situation the above eight haplotypes have the equal probability of occurrence 1/8 in individuals. As a result, the haplotype population reaches the highest diversity as the population entropy attains the maximum value of $\log 8$ (Jones, Chapter 2, 1979). Now we consider the situation where there exist some dependences among the three loci. Note that these dependences are apparent as increased frequencies of specific haplotypes compared to what would be expected if alleles at the three loci are combined at random. For example, if we set $p(h_1) = 1/2$, $p(h_5) = 1/2$, and $p(h_i) = 0$, $i \neq 1, 5$, then the three loci are fully determined by the first locus. With the entropy being equal to $\log 2$, the resulting haplotype population yields a smaller diversity than the previous one. We observe that as an empirical version of the population entropy, $S(\hat{\boldsymbol{H}})$ is close to its population value when n is large. Therefore, in general the population haplotype entropy, and thus its empirical version $S(\hat{\boldsymbol{H}})$, tends to decrease as the strength of these dependences increases.

Testing for interaction between two haplotype blocks

Let $\boldsymbol{G} = (G_1, \ldots, G_n)^T$ be partitioned into two blocks, say

$$\boldsymbol{G} = (\boldsymbol{G}^{(1)}, \boldsymbol{G}^{(2)}) = \left(\begin{pmatrix} G_1^{(1)} \\ G_1^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} G_n^{(1)} \\ G_n^{(2)} \end{pmatrix} \right)^T .$$

Suppose we are interested in testing if there exists interaction between the two blocks $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$. This problem can be stated as testing the hypotheses: the two blocks are independent (null hypothesis) versus the two blocks are dependent (alternative hypothesis). As pointed out in the previous section, if the null hypothesis is true, $s(\boldsymbol{G})$ will tend to have a large value, otherwise it will tend to be small. Hence, $s(\boldsymbol{G})$ can be used as a test statistic for this test. Because the distribution of $s(\boldsymbol{G})$ under the null hypothesis is unknown, the following procedure is designed to calculate the p-values of the test.

(1) Generate n' random permutations of $(G_1^{(2)}, \ldots, G_n^{(2)})$, and denote them by $(G_{j,1}^{(2)}, \ldots, G_{j,n}^{(2)})$, $j = 1, \ldots, n'$.

(2) Form a random sample $\boldsymbol{G}_j^*$, $j = 1, \ldots, n'$, where $\boldsymbol{G}_j^*$ is formed by pairing $(G_1^{(1)}, \ldots, G_n^{(1)})$ with $(G_{j,1}^{(2)}, \ldots, G_{j,n}^{(2)})$.

(3) Calculate the haplotype entropy for each $\boldsymbol{G}_j^*$. An empirical p-value can then be defined by the proportion of $s(\boldsymbol{G}_j^*)$'s that are less than or equal to $s(\boldsymbol{G})$, i.e., $\#\{s(\boldsymbol{G}_j^*) : \ s(\boldsymbol{G}_j^*) \leq s(\boldsymbol{G})\}/n'$.

The number n' is usually set to a moderate number. For example, it is 500 and 1000 in this paper.

Based on the central limit theorem, an empirical z-score statistic,

$$Z(\boldsymbol{G}) = \frac{s(\boldsymbol{G}) - A}{\sqrt{V}}$$

can also be defined for the test, where A and V are the sample mean and variance of the $s(\boldsymbol{G}_j^*)$'s. The empirical p-value calculated in step (3) can be used to examine whether the between block interaction existing in $\boldsymbol{G}$ was obtained by chance or not, whereas the empirical z-score statistic more sensitively measures how large the distance of the genotypes under investigation is from the population of genotypes without block interactions.

The above procedure will be used below to test the significance of the pairwise interactions among haplotype blocks or loci. In each case the significance of an interaction will be decided by a threshold for p-values. Assessment of the overall significance to account for multiple testing, is not straightforward because there are many correlations among the tests. An alternative approach is to control the false discovery rate (FDR), which is defined by the expected proportion of false positives among those called significant: $E[V^*/R^*|R^* > 0]$. Here for a given threshold, V^* is the total number of false positives while R^* is the total number of interactions called significant according the threshold. We opt for the recent proposal of Storey and Tibshirani (2003) to estimate the FDR and calculate the q-value, a measure of statistical significance in terms of FDR, for each individual test under dependence.

Structure annealing algorithm

In this section, we propose a new algorithm, the so-called structure annealing algorithm, to minimize $S(\boldsymbol{H})$. The algorithm is proposed based on the following observation. Let $\boldsymbol{G} = (\boldsymbol{G}^{(1)}, \boldsymbol{G}^{(2)})$ be a random partition of $\boldsymbol{G}$, and $\boldsymbol{H} = (\boldsymbol{H}^{(1)}, \boldsymbol{H}^{(2)})$ be the corresponding partition of $\boldsymbol{H}$. It is easy to see that if $\boldsymbol{H}$ is compatible with $\boldsymbol{G}$, then $\boldsymbol{H}^{(1)}$ is compatible with $G^{(1)}$. Furthermore, if $S(\boldsymbol{H}^{(1)})$ is a good approximation to $s(\boldsymbol{G}^{(1)})$, then $S(\boldsymbol{H}^{(1)}, \boldsymbol{H}^{(2)})$ should be a good approximation to $s(\boldsymbol{G})$, provided that $\boldsymbol{H}$ is compatible with $\boldsymbol{G}$ and the number of loci in $\boldsymbol{G}^{(2)}$ is not large. This observation motivates the following sequential way to minimize the objective function $S(\boldsymbol{H})$.

Suppose now that $\boldsymbol{G}$ is partitioned into z blocks, $\boldsymbol{G} = (\boldsymbol{G}^{(1)}, \ldots, \boldsymbol{G}^{(z)})$, where $\boldsymbol{G}^{(b)}$ comprises k_b loci and $\sum_{b=1}^{z} k_b = L$. Preferably, k_b is set to a small number, for example, $k_b \leq 8$ for all examples of this paper. The structure annealing algorithm consists of two building blocks: a local updating algorithm and an extrapolation algorithm. The local updating algorithm (described in Appendix A) is designed to simulate from the distributions

$$P(\tilde{\boldsymbol{H}}^{(b)}) \propto \exp\{-S(\tilde{\boldsymbol{H}}^{(b)})/t_b\},$$

for $b = 1, \ldots, z$, where t_b is called the temperature of this distribution, and $\tilde{\boldsymbol{H}}^{(b)} = (\boldsymbol{H}^{(1)}, \ldots, \boldsymbol{H}^{(b)})$ which is compatible with $\tilde{G}^{(b)} = (\boldsymbol{G}^{(1)}, \ldots, \boldsymbol{G}^{(b)})$. The extrapolation algorithm (described in Appendix B) is designed to extrapolate $\tilde{\boldsymbol{H}}^{(b)}$ to $\tilde{\boldsymbol{H}}^{(b+1)}$. The structure annealing algorithm starts with the simulation from $P(\tilde{\boldsymbol{H}}^{(1)})$ by the local updating algorithm, where $\tilde{\boldsymbol{H}}^{(1)} = (\boldsymbol{H}^{(1)})$. Since the block size of $\boldsymbol{G}^{(1)}$ is usually small, the iteration number of the local updating steps is also moderate at this step. We denote this iteration number by m_1, and set $m_1 = 10000$ for all examples of this paper. Then the algorithm proceeds for $z - 1$ steps. The $(b+1)^{th}$ step consists of two sub-steps which are described as follows.

(a) (Extrapolation) Extrapolate the haplotype $\tilde{\boldsymbol{H}}^{(b)}$, which is obtained at the last iteration of the b^{th} step, to a compatible haplotype pair of $\tilde{\boldsymbol{G}}^{(b+1)}$.

(b) (Local updating) Simulate from the distribution $P(\tilde{\boldsymbol{H}}^{(b+1)})$ by the local updating algorithm for m_{b+1} steps.

The m_b is a monotone increasing function of b, for example, we set $m_b = m_1 \times b$ for $b = 1, 2, \ldots, z-1$ and $m_z = 10 \times m_1 \times z$. Here we follow simulated annealing (Kirkpatrick et al. 1983) to set a large iteration number for the last step simulation.

Results

Simulated data sets

We used a coalescent-based program of Professor R. Hudson, named **MS**, to simulate haplotypes for four different situations described by quantities $(\theta, R) = (4, 0), (4, 4), (4, 20)$ and $(16, 16)$ respectively. Here $\theta = 4N_e\mu$, $R = 4N_e r$, N_e is the effective population size, μ is the total per-generation mutation rate across the region sequenced, and r is the length, in Morgans, of the region sequenced. For each setting of (θ, R), this generated 40 independent data sets, each containing 40 haplotypes. For each data set the haplotypes were randomly paired to form 20 genotypes. As a result, for each case of (θ, R), we had 40 sets of 20 genotypes. They are denoted by $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_{40}$, with $\boldsymbol{G}_i = (G_{i,1}, \ldots, G_{i,20})^T$. We split each of $G_{i,j}$'s into two parts, $G_{i,j}^{(1)}$ and $G_{i,j}^{(2)}$, of equal length for $i = 1, \ldots, 40$ and $j = 1, \ldots, 20$. Totally we have 80 genotype segments. With these segments, 20 new data sets, which are denoted by $\boldsymbol{G}_1^*, \ldots, \boldsymbol{G}_{20}^*$, are formed, where $\boldsymbol{G}_k^*$ is formed by attaching the segment $G_{20+k,j}^{(2)}$ to the segment $G_{k,j}^{(1)}$ for $k = 1, \ldots, 20$. The above construction procedure shows that there exist two independent blocks in each of $\boldsymbol{G}_k^*$'s.

In the following we will regard $\boldsymbol{G}_1^*, \ldots, \boldsymbol{G}_{20}^*$ as samples from a population of which the two genotype blocks are independent, while we will regard $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_{20}$ as samples from a population of which the two genotype blocks are dependent. To evaluate the power of our procedure, we applied it to these genotype data sets. The resulting p-values and z-scores were summarised in Figure 1. To find the interesting blocks, we further analysed these p-values by setting the lower and upper thresholds of 0.01 and 0.15. We say two blocks are dependent if the corresponding p-value is less than or equal to 0.01, whereas we say they are independent if the corresponding p-value is larger than or equal to 0.15. The performance of our procedure is measured by the proportions of false positives and negatives, F_a and F_n. That is, F_a is the proportion of falsely rejecting the null hypothesis when the null hypothesis is true, and F_n is the proportion of falsely not rejecting the null hypothesis when the alternative is true. For the above simulated data, we have $(F_a, F_n) = (0, 2/20)$ when $(\theta, R) = (4, 0)$; and $(F_a, F_n) = (0, 0)$ when $(\theta, R) = (4, 4)$, (4,20), and (16,16). These results show that our procedure is indeed an effective tool for detecting haplotype-interactions. As pointed out in the Introduction, the coalescent model can capture certain main features in a haplotype population (Stephens et al. 2001). The above simulated coalescent models might share some common features with real haplotype data. Thus these thresholds were used to guide our choice of the corresponding thresholds when we applied our method to the Dutch type 1 diabetes data below.

Put Figure 1 here.

Type 1 diabetes data

36 candidate genes, listed in Table 1, were selected from the pathways, which are potentially implicated in the development and progression of atherosclerosis: lipid and homocysteine metabolism, regulation of blood pressure and coagulation, inflammation, cellular adhesion, and matrix integrity (Cheng et al. 1999; Dassen et al. 2001). All of them have been reported in the database called Online Mendelian Inheritance in Man (OMIM). Then Dassen et al. (2001) described an assay for genotyping a panel of 65 SNPs

that represent variation within these genes, which is an early version of RMS Research Assay for cardiovascular disease (CVD) Genetics designed by Roche Molecular Systems, Inc. Most of these SNPs have been shown to be implicated with some metabolic diseases such as cardiovascular disease, coronary artery disease, hypertension, asthma, obesity, atherosclerosis, myocardial infarction, hyperlipidemia, Alzheimer disease, and so on. See Table 1 and the database OMIM for more details. The rest of these SNPs are either the polymorphisms at (or close to) the promoter regions that may (directly or indirectly) play certain dysregulation role for the genes of interest or the polymorphisms at coding regions with nonsynonymous changes (Dassen et al. 2001; Cheng et al. 1999; Flori et al. 2003; Vatay et al. 2003). For example, V67 was selected because it could have a protective role against type 2 diabetes (NIDDM) (Vatay et al. 2003). V66 was included as it often interfered with our ability to call V67 correctly. We had no prior functional information, other than that its proximity to V67 could mean that it would also have impact on the function of the gene TNF. As pointed out in the Introduction, since both diabetes mellitus and atherosclerosis can be regarded as metabolic diseases, with many overlapping biochemical and clinical parameters, the variants that are susceptible to atherosclerosis may also be the cause of type 1 diabetes. So this assay was also applied to a Dutch diabetes cohort, which includes 136 unrelated individuals (89 type 1 diabetes patients with impaired endothelial function, and 47 healthy controls). Endothelial function was assessed by measuring changes in forearm blood flow after pharmacological interventions. The DNA samples from the 136 individuals were genotyped by using the polymerase chain reaction (PCR). This led to 136 genotypes of 65 loci. 9 loci (V58, V59, V66, V67, V5, V57, V51, V52, V30) were not used in our following data analysis since these loci have the so-called heavy missing problem, where at least 21% of the 136 individual genotypes were incomplete in the PCR experiments. The heavy missing may introduce the bias in our data analysis. The cutting-point 21% was selected according to our experience. We ended up with a 136 by 56 data matrix. Each genotype can be divided into 16 blocks according to their chromosome identities. See Table 1 for more details.

Put Table 1 here.

We started with the search for pair-wise interactions among these 16 unlinked blocks. The search was performed on the cases and controls separately. The p-values for the cases and controls were contrasted by plotting them in Figures 2 and 3, respectively.

Put Figures 2 and 3 here.

We obtained 10 pairs of interacting blocks, which are located on chromosome pairs (1,4), (1,12), (1,15), (1,19), (2,7), (3,19), (5,7), (6,7), (6,21), (7,11), and (17,21), respectively. See Table 2 for more details. These block pairs were selected by the following criteria: for the up-interaction, we claimed that there is an increase in haplotype-interaction if the p-value of the controls is larger than 0.15, and the p-value of the cases is less than or equal to 0.01, and the z-score of the cases is less than or equal to -2. This says that in contrast to the healthy individuals, there is a significant interaction between two haplotype-blocks under consideration in the disease individuals. For the down-interaction, we claimed that there is a decrease in haplotype-interaction if the p-value of the cases is larger than 0.15, and the p-value of the controls is less than or equal to 0.01, and the z-score of the controls is less than or equal to -2. This implies that in contrast to the healthy individuals, there is no significant interaction between two haplotype-blocks under consideration in the disease individuals. Among these selected blocks, the up-interaction pairs for chromosome pairs (1,4), (1,12), (1,15), (1,19), (6,7), and (17,21) indicate that the pathways harboring these variants may have been modified by adding some interactions between some genes due to the disease. Analogously, the down-interaction pairs on chromosome pairs (3,19), (2,7), (6,21), (7,11), (5,7), (12,15) indicate that the related pathways may have been changed as interactions between some genes are disrupted. Note with the p-value thresholds 0.01 and 0.15 for cases and controls respectively, the corresponding estimated FDRs of these multiple tests for cases and controls are 0.017 and 0.029. There will be more interaction pairs if we take 0.035 and 0.2 as the thresholds for cases and controls, respectively. The FDRs will then become 0.040 and 0.048. All the p-values are below 0.05. See Table 2 for more details.

To see how these interactions modify the related pathways, we ran our procedure on the pairs of variants on these blocks. Consequently 25 pairs of variants were found to show certain evidence of susceptibility to the disease. Table 2 indicates that these variants are distributed on 19 genes: NPPA, SELE, ADOB, AGTR1 ADRB2, LPA, TNF, TNFb, DCP1, ADD1, SCNN1A, APOE, NOS3, LPL, LIPC, PON1, CBS, APOA4, and APOC3. Note that APOB, ADRB2, LPA, APOE, LPL, LIPC, PON1, and APOA4 are on the pathway of lipid metabolism; CBS is on the pathway of homocysteine metabolism; NPPA, AGTR1, ADRB2, DCP1, SCNN1A, NOS3 are on the pathway of blood pressure; SELE is on the pathway of coagulation; SELE, TNF, TNFb are on the pathway of inflammation; and ADD1 are on the pathway of matrix integrity. Thus, within the pathway of lipid metabolism there are seven up- or down-interactions, denoted by the symbols (+) and (-) respectively, among some genes. They are V9:V22 (APOB:LPL) (+), V8:V20 (APOB:LIPC) (-), V4:V26 (LPA:PON1) (+), V26:V7 (PON1:APOA4) (-), V25:V10 (PON1:APOC3) (-) and V25:V12 (PON1:APOC3) (-). These interactions are predisposing to the disease. Similarly, within the pathway of blood pressure there is one down-interaction: V50:V38 (ADRB2:NOS3) (-). The rest are related to interactions among the six pathways mentioned above. Here up-interaction (down-interaction) is trying to capture the biological phenomenon that the pathways of lipid metabolism, homocysteine metabolism, blood pressure, inflammation, and matrix integrity are modified by creating (disrupting) interactions among some genes that lie in these pathways. Similar to Sudbery (p.144, 1998), the up-interactions would suggest that those interactions lead to a susceptibility to the disease while the down-interactions could imply that the related interactions may have a protective effect on developing the disease. These results indicate a complicated feature of (possibly non-multiplicative) effects of the interactions on the risk for type 1 diabetes.

Note that Dassen et al. (2001) have identified a set of dominant variants: V4, V15, V28, and V50, which are on chromosomes 6, 11, 19, and 5. This combined with the above results yields the following transitive and disease-predisposing variants in the sense that there are significant increases (or decreases) of interactions of these variants with some dominant variants: V26, V37, V38, V39, V7, V8, V10, V11, V12, V13, V65, V68, V20, V25, and V47.

Put Table 2 here.

In the next step, we screened for interactions in linked regions. For simplicity, we adopted the following strategy. Taking Block 1 as example, we sequentially tested 6 sub-block pairs for the cases and controls: the first pair $\{1\}, \{2, 3, 4, 5, 7\}$ with 1 being the splitting-location; the second pair $\{1, 2\}, \{3, 4, 5, 7\}$ with 2 being the splitting-location; and so on. Here the numbers 1, 2, 3, 4, 5, 6, and 7 denote 7 variants in Block 1. The six sub-block pairs are uniquely defined by six splitting-locations 1, 2, 3, 4, 5, and 6. We compared the resulting six pairs of p-values and z-scores in Table 3. It suggests that there exists some disease predisposing interaction between sub-block pairs $\{1, 2, 3, 4\}$ and $\{5, 6, 7\}$. Following the same argument as above, for block 6 we may conclude that variant V64 might be a transitive disease-predisposing variant because the dominant variant V4 is at sub-block $\{1, 2, 3\}$. The evidence of disease-predisposing interactions within the other blocks are reported in Table 4, which yields the transitive variant V14. Note that in practice we need to test the interactions for all bi-partitions of seven loci since the strength of linkage disequilibrium patterns is not, typically, a monotonic function of genetic distance. Our procedure can be easily extended to this general setting as it does not use any information on genetic distances among these loci.

Put Tables 3 and 4 here.

Discussion

The logistic regression mentioned in the Introduction is a very important genotype-based tool for detecting dominant polymorphisms and epistatic effects (i.e., genotype interactions) that are associated with the disease. One disadvantage of this method over some haplotype-based methods is that it ignores the potential disease-predisposing haplotype interactions. To contend this disadvantage, we have presented a procedure for evaluating the contributions of these haplotype-interactions to susceptibility of disease, in which the entropy is used to measure the diversity of a haplotype population. Our procedure can be easily generalized to other measures of the haplotype-diversity (Clayton 2002; Weir 1996). Of course, for applications, we should combine these two methods together in order to extract more complete information from unphased genotype data in the following steps: first, apply the logistic regression to detect dominant disease-predisposing variants and genotype-interactions. Then, as a complement, use our procedure to find potential haplotype-interactions. Finally, the transitive variants are predicted by finding these variants that are interacting with the dominant ones.

In the first step, we assume to have a sample of n_1 cases and n_2 controls, each is genotyped at m polymorphisms. Let p_j be the probability of individual j being a case rather than a control. Following McCullagh and Nelder (1989), we model p_j as

$$\mathrm{logit}(p_j) = \log(\frac{p_j}{1-p_j}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$$

where $x_1, \ldots, x_m$ are covariates depending on the genotypes of the individual and $\beta_0, \ldots, \beta_m$ are coefficient to be estimated. To examine the effects of a set of polymorphisms, we can test whether the data is significantly better represented when these polymorphisms are included in the model compared to when they are not in the model, using likelihood ratio tests (Cordell and Clayton 2002). This is equivalent to testing whether the corresponding coefficients are significantly different from 0. Similarly, we can account for the genotype interactions by adding some epistatic terms to the above model. A commonly used strategy for evaluation of the effects of the different polymorphisms is to fit these models in a stepwise fashion. Following Cordell and Clayton (2002), for the Dutch type 1 diabetes data, we first code $x_j = -0.5, 0.5, 0.5$ for genotypes $0, 2, 1$, and also code 0.5 for the cases where genotypes are missing. We set 0.05 as a nominal significance level for all these tests involved in the stepwise logistic-regression procedure. This yields seven dominant disease susceptible alleles on chromosomes 3, 6, 7, 6, 11, 2, 19 respectively: V41(AA), V4(TT), V26(GG), V64(GG), V15(GG), V28(–), V9(missing), and one genotype interaction between V41(AA) and V64(GG), where for example in the notation V41(AA), V41 is the name of the variant while (AA) is one of its alleles. See Table 5. The result is slightly different from the prediction-based logistic-regression procedure of Dassen et al. (2001). This might be due to different criteria being used.

Put Table 5 here.

In the second step, we start with search for the haplotype interaction between blocks located on different chromosomes followed by testing the interactions within each blocks. If two blocks are found interacting, we can further narrow the search area to identify which variants in the blocks are involved in this interaction. For the Dutch type 1 diabetes data, in the previous section we have shown 9 pairs of interacting blocks, which are predisposed to type 1 diabetes. Combining with the result from the first step, we can infer some transitive disease-predisposing variants as shown in the previous section. The results demonstrate a complicated gene-gene interaction network, which might predispose to type 1 diabetes through modifying the pathways of lipid metabolism, blood pressure, inflammation, coagulation, and matrix integrity.

Use of interaction between unlinked genomic regions has been suggested for improving power of detecting loci of small effects on the disease phenotype, for example, in type 1 diabetes (Bugawan et al. 2003; Cordell et al. 1995, 2000), type 2 diabetes (Cox et al. 1999) and inflammatory bowel disease (Cho et al. 1998). Cordell et al. (1995) reported that there are interactions between the loci IDDM1 (chromosome 6p21) and IDDM2 (chromosome 11p15) and between the loci IDDM1 and IDDM4 (on chromosome 11q13.3) in the context of the logistic regression model. Cox et al. (1999) showed that the loci on chromosomes 2 and 15 interact to increase susceptibility to type 2 diabetes in the context of nonparametric lod score. Cox et al. (2001) made a systematic screen for correlation between family-specific non-parametric lod scores in order to evaluate evidence of interactions between some unlinked regions on chromosomes 1, 2, 3, 4, 6, 11, and 19. These methods are usually restricted to family data. Unlike these authors, we focus here on interactions between genetic variants in a list of potential candidate genes across a number of chromosomes, where some of these variants have already been shown to be associated with some metabolic diseases. Moreover, the proposed approach is specified for unphased genotype data (possibly with missing problems) from case-control studies. Thus, our method could be a valuable contribution to a genome-wide association study of a complex disease, especially when direct determination of the molecular haplotypes from experiment or family data is not feasible.

Although significant and consistent linkage evidence was reported for the susceptibility intervals IDDM8 (on chromosome 6q27), IDDM4 (on 11q) and IDDM5 (on 6q25), evidence for most other intervals varies in different data sets—probably due to a weak effect of the disease genes, genetic heterogeneity, random variation or inappropriate correction for multiple tests (see Pugliese 2001). To reduce the possible effect of genetic heterogeneity, we need to confirm our initial finding by analysing other populations in future studies. Since we compared correlated variants, it is important to take into account the potential effects of multiple tests on the power of our procedure. For our case, there are 120 pairwise tests among 16 haplotype blocks. A simple Bonferroni (or Dunn-Sidak) correction leads to the adjusted threshold of 4.17×10^{-4} for p-values if we want to achieve the significant level of 0.05. Then there are only 7 block-pairs in Table 2 remained nominally significant after this correction. Such a correction seems too conservative due to high dependences among these tests. This has been confirmed by Bugawan et al. (2003) based on a permutation procedure. Unfortunately using resampling methods such as permutation can be computationally prohibitive in our case. However, we have shown that the recently developed procedure of Storey and Tibshirani (2003) is applicable to our setting.

Acknowledgements We are grateful the Editor, the Deputy Editor, and other members in the Editorial Board, and two anonymous reviewers for their very constructive comments that have led to improvement of the presentation and results of the paper. We thank Roche Molecular Systems, Inc. (Alameda, CA) for providing the multiplex genotyping assays under a research collaboration. We thank Dr. Suzanne Cheng for kindly providing some information on the SNPs used in this paper. We also thank Professor W. van Zwet and Dr. P. Lindsey for useful discussions, and Wim J. M. Senden for computing supports. This work was partially done when the first author was working in EURANDOM, The Netherlands. The work was supported in part by the programme of computational molecular biology in EURANDOM, by the Institute for Mathematical Sciences, National University of Singapore, and by a grant (No. 01/1/21/19/217) from BMRC-NSTB of Singapore.

Appendix A: Local updating

The local updating algorithm includes two operators, ν-mutation and peer learning. In every iteration, they are selected to perform with probability 0.2 and 0.8, respectively. Of course, the probabilities can be tuned by the user. But a large performing probability is usually assigned to the peer learning operator, as it tends to force the haplotypes to coalesce. The two operators are described as follows.

ν-mutation

In the ν-mutation operator, a total of $\max\{1, \gamma_b \nu\}$ haplotype pairs at the heterozygous ($g_{ij} = 2$) or missing loci ($g_{ij} = 9, 8, 7$) are randomly selected to undergo changes, where γ_b is the total number of heterozygous and missing loci in $\tilde{\boldsymbol{G}}_b$, and ν is the mutation rate specified by the user. The ν is usually set to a small number, for example, we set $\nu = 0.001$ for all examples of this paper. The changes are accepted or rejected according to the Metropolis-Hastings rule (Metropolis et al. 1953; Hastings 1970), i.e., the new haplotypes $\tilde{\boldsymbol{H}}_*^{(b)}$ are accepted with probability $\min(1, r_m)$, where

$$r_m = \exp\{-[s(\tilde{\boldsymbol{H}}_*^{(b)}) - s(\tilde{\boldsymbol{H}}^{(b)})]/t_b\} \frac{T(\tilde{\boldsymbol{H}}_*^{(b)} \to \tilde{\boldsymbol{H}}^{(b)})}{T(\tilde{\boldsymbol{H}}^{(b)} \to \tilde{\boldsymbol{H}}_*^{(b)})},$$

where $T(\cdot \to \cdot)$ denotes the transition probability between the current and new haplotypes. The transition proceeds as follows. If the pair $(h_{b,ij,1}, h_{b,ij,2})$ is selected to undergo a change, and if $g_{ij} = 2$, then the values of $h_{b,ij,1}$ and $h_{b,ij,2}$ will be simply swapped by setting $h_{b,ij,1} = 1 - h_{b,ij,1}$ and $h_{b,ij,2} = 1 - h_{b,ij,2}$. If $g_{ij} = 9$, one of the pairs $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ will be re-assigned to $(h_{b,ij,1}, h_{b,ij,2})$ equally likely. Similarly, if $g_{ij} = 8$ or 7, one of the possible haplotype pairs will also be re-assigned to $(h_{b,ij,1}, h_{b,ij,2})$ equally likely. The other selected haplotype pairs will be mutated in the same way, but independently. It is easy to see that transition is symmetric in the sense that $T(\tilde{\boldsymbol{H}}_*^{(b)} \to \tilde{\boldsymbol{H}}^{(b)}) = T(\tilde{\boldsymbol{H}}^{(b)} \to \tilde{\boldsymbol{H}}_*^{(b)})$.

Peer learning

The peer learning operator works as follows.

(a) Randomly select one haplotype, say $h_{b,u,v}$, from the set $\{h_{b,1,1}, h_{b,1,2}; \ldots ; h_{b,n,1}, h_{b,n,2}\}$.

(b) Randomly select one haplotype, say $h_{b,s,t}$, from the set $\{h_{b,1,1}, h_{b,1,2}; \ldots ;h_{b,u-1,1}, h_{b,u-1,2}; h_{b,u+1,1}, h_{b,u+1,2}; \ldots ; h_{b,n,1}, h_{b,n,2}\}$ with probability $w_{b,s,t}/\sum_{i\neq u}\sum_{j=1}^{2} w_{b,i,j}$, where $w_{b,i,j} = \exp\{-d(h_{b,u,v}, h_{b,i,j})/t_{\text{sel}}\}$, $d(h_{b,u,v}, h_{b,i,j})$ is the number of different haplotypes at the first $\sum_{i=1}^{b} k_i$ loci of $h_{b,u,v}$ and $h_{b,i,j}$, and t_{sel} is the so-called selection temperature.

(c) For each genotype g_{uj}, if $g_{uj} = 0$ or 1, we keep $h_{b,uj,v}$ unchanged; if $g_{uj} = 2$, 9, 8, or 7 and $h_{b,uj,v} = h_{b,sj,t}$, we keep $h_{b,uj,v}$ unchanged with probability p_l, and change $h_{b,uj,v}$ to $h_{b,sj,t}$ with probability $1 - p_l$; if $g_{uj} = 2$, 9, 8, or 7 and $h_{b,uj,(v)} \neq h_{b,sj,t}$, we keep $h_{b,uj,v}$ unchanged with probability $1 - p_l$, and change $h_{b,uj,v}$ to $h_{b,sj,t}$ with probability p_l. Update the complementary pair of $h_{b,u,v}$ accordingly such that they are compatible with g_u.

(d) According to the Metropolis-Hastings rule, accept the new haplotype pair with probability $\min\{1, r_l\}$, where

$$r_l = \exp\{-[s(\tilde{\boldsymbol{H}}_*^{(b))} - s(\tilde{\boldsymbol{H}}^{(b)})]/t_b\}\frac{T(\tilde{\boldsymbol{H}}_*^{(b)} \to \tilde{\boldsymbol{H}}^{(b)})}{T(\tilde{\boldsymbol{H}}^{(b)} \to \tilde{\boldsymbol{H}}_*^{(b)})}.$$

Here the transition probability equals

$$T(\tilde{\boldsymbol{H}}^{(b)} \to \tilde{\boldsymbol{H}}_*^{(b)}) = p_l^{\alpha_1}(1-p_l)^{\alpha_2}(\frac{1}{2})^{\alpha_3},$$

where α_1 is the total number of the common haplotypes of $\tilde{\boldsymbol{H}}_*^{(b)}$ and $\tilde{\boldsymbol{H}}^{(b)}$ at the heterozygous and missing loci; α_2 is the total number of the different haplotypes of $\tilde{\boldsymbol{H}}_*^{(b)}$ and $\tilde{\boldsymbol{H}}^{(b)}$ at the heterozygous and missing loci; and α_3 counts the total number of times of randomly assigning the haplotype values in the complementary haplotype pair of $h_{b,u,v}$. The p_l is a user-specified parameter. We set $p_l = 0.9$ for all examples of this paper. The transition probability $T(\tilde{\boldsymbol{H}}_*^{(b)} \rightarrow \tilde{\boldsymbol{H}}^{(b)})$ can be computed similarly.

This operator makes it possible for haplotypes to coalesce together very fast if it is feasible.

Appendix B: Extrapolation

The extrapolation operator extrapolates $\tilde{\boldsymbol{H}}^{(b)}$ to $\tilde{\boldsymbol{H}}^{(b+1)}$ by attaching the haplotype pairs compatible with $\boldsymbol{G}^{(b+1)}$. Let's call a haplotype "original" if it first appears in $\tilde{\boldsymbol{H}}^{(b)}$ in some scanning order, for instance, the natural order $(h_{b,1,1}, h_{b,1,2}; \ldots ; h_{b,n,1}, h_{b,n,2})$ used in this paper, where $(h_{b,i,1}, h_{b,i,2})$ is the haplotype pair of the i-th genotype in $\tilde{\boldsymbol{G}}^{(b)}$; otherwise, we call it "duplicate". The extrapolation proceeds in the pre-fixed scanning order as follows. If a haplotype and its complementary pair are both "original", it is extrapolated independently, i.e., if g_{ij} is a heterozygous or missing allele, then $(h_{b+1,ij,1}, h_{b+1,ij,2})$ is equally likely set to one of the possible haplotype pairs. If a haplotype is "duplicate", then it will be extrapolated according to the corresponding original copy. Note that in this case the extrapolation for the corresponding original copy has been finished. For example, if $h_{b,u,v}$ is a duplicate of $h_{b,s,t}$, if g_{uj} is a heterozygous or missing allele, then $h_{b+1,uj,v}$ will be set to the same value as $h_{b+1,sj,t}$ with probability p_e, and a different value from $h_{b+1,sj,t}$ with probability $1 - p_e$. The complementary pair of $h_{b+1,uj,v}$ will be set accordingly such that the pair is compatible with g_u. We usually set p_e to a large value, say, 0.95, for all examples of this paper. Obviously, the extrapolation operator will provide a good starting point for the simulation from the distribution $P(\tilde{\boldsymbol{H}}^{(b+1)})$.

Electronic-Database Information

URLs for data in this article are as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nih.gov/Omin/ (for the 36 candidate genes in Table 1).

Online SNP database, http://www.ncbi.nih.gov/SNP/ (for the 65 SNPs in Table 1).

http://research.bwh.harvard.edu/ca16ref.doc (for more references to these 65 SNPs).

References

Bugawan TL, Mirel DB, Valdes AM, Panelo A, Pozzilli P, Erlich HA (2003) Association and interaction of the IL4R, IL4, and IL13 loci with type 1 diabetes among Filipinos. Am J Hum Genet, 72:1505-1514

Cheng S, Grow MA, Pallaud C, Klitz W, Erlich HA, Visvikis S, Chen JJ, Pullinger CR, Malloy MJ, Siest G, Kane JP (1999) A multilocus genotyping assay for candidate markers of cardiovascular disease risk. Genome Res 9:936-949

Cho JH, Nicolae DL, Gold CT, Fields MC, Labuda MC, Rohal PM, Pickles MR, Qin L, Fu Y, Mann JS, Kirschner BS, Jabs EW, Weber J, Hanauer SB, Bayless TM, Brant SR (1998) Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistas is between 1p and IBD1. Proc Natl Acad Sci USA 95:7502-7507

Clayton D (2002) Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/ htsnp.pdf (accessed September 5, 2003)

Cordell HJ, Clayton DG (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70:124-141

Cordell HJ, Todd JA, Bennett ST, Kawagushi Y, Farrall M (1995) Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. Am J Hum Genet 57:920-934

Cordell J, Todd JA (1995) Multifactorial inheritance in type 1 diabetes. TIG 11:499-504

Cordell J, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB, Wicker LS, Clayton DG (2001) Statistical modelling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. Genetics 158:357-367

Cordell HJ, Wedig GC, Jacobs KB, Elston RC (2000) Multilocus linkage tests based on affected relative pairs. Am J Hum Genet. 66:1273-1286

Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A (1999) Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nat Genet 21:213-215

Cox NJ, Wapelhorst B, Morrison VA, Johnson L, Pinchuk L, Spielman RS, Todd JA, Concannon P (2001) Seven regions of the genome show evidence of linkage to type 1 diabetes in a consensus analysis of 767 multiplex families. Am J Hum Genet 69:820-830

Cucca F, Dudbridge F, Loddo M, Mulargia AP, Lampis R, Angius E, De Virgiliis, Koeleman BP, Bain SC, Barnett AH, Gilchrist F, Cordell H, Welsh K, Todd JA (2001) The HLA-DPB1-associated component of the IDDM1 and its relationship to the major loci HLA-DQB1, -DQA1, and -DRB1. Diabetes 50:1200-1205

Dassen W, Spiering W, de Leeuw P, Smits P, Dijk WA, Spruijt H, Gommer E., Bonnemayer C, Doevendans PA (2001) Unravelling gene interactions to find the cause of artherosclerosis, a multigenic disease, using an artificial neural network. Computers in Cardiology, 28:373-376

Drysdale CM, McGraw DW, Stack CB, Stephens JC Judson RS, Nadabalan K, Arnold K, Ruano G, Liggett SB (2000) Complex promoter and coding region β_2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci USA 97:10483-10488

Fain PR, Eisenbarth GS (2001) Type 1 diabetes, autoimmunity and the MHC. In: Lowe WL Jr., editor. Genetics of Diabetes Mellitus. Kluwer Academic Publishers, Boston. p 43-64

Flori L, Sawadogo S, Esnault C, Delahaye NF, Fumoux F, Rihet P (2003) Linkage of mild malaria to the major histocompatibility complex in families living in Burkina Faso. Hum Mol Genet 12:375-378

Friday RP, Trucco M, Pietropaolo M (1999) Genetics of type 1 diabetes mellitus. Diabetes Nutr Metab, 12:3-26

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109

Kirkpatrick, S, Gelatt, Jr, CD, Vecchi, MP (1983) Optimization by simulated annealing. Science, 220: 671-680

Koeleman BP, Dudbridge F, Cordell HJ, Todd JA (2002) Adaptation of the extended transmission/disequilibrium test to distinguish disease associations of multiple loci: the conditional extended transmission/disequilibrium test. Ann Hum Genet 64:207-213

Jones DS (1979) Elementary Information Theory. Clarendon Press, Oxford

McCullach P, Nelder JA (1989) Generalized Linear Models, 2nd Edition. Chapman & Hall, London

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machine. J Chem Phys 21:1087-1091

Noble JA, Valdes AM, Bugawan TL, Apple RJ, Thomson G, Erlich HA. (2002) The HLA class I A locus affects susceptibility to type 1 diabetes. Hum Immunol 63:657-64

Pugliese A (2001) Genetic factors in type 1 diabetes. In: Lowe WL Jr., editor. Genetics of Diabetes Mellitus. Kluwer Academic Publishers, Boston. p 25-42

Schranz DB, Lermark A (1998) Immunology in diabetes: an update. Diabetes-Metabolism Rev 14:3-29

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

Storey JD, Tibshirani R (2003) Statistical significance for genome-wide experiments. Proc Natl Acad Sci USA (in press)

Sudbery P (1998) Human Molecular Genetics. Addison Wesley Longman, Harlow

Valdes AM, Thomson G (1997) Detecting disease-predisposing variants: the haplotype method. Am J Hum Genet 60:703-716

Vatay A, Yang Y, Chung EK, Zhou B, Blanchong CA, Kovács M, Kard́i I, Füst G, Romics L, Varga L, Yu Y, Szalai C (2003) Relationship between complement components C4A and C4B diversities and two TNFA promoter polymorphisms in two healthy Caucasian populations. Hum Immunol 64:543-552.

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in human genome. Nat Rev 4:587-597

Weir BS (1996) Genetic Data Analysis II. Sinauer Associates, Massachusetts

Zhang J, Liang F, Hoehe M, Vingron M (2001) On haplotype reconstruction for diploid organisms. EURANDOM Report-2001-026

Table 1: SNPs used in this study

block	variant (symbol, dbSNPrs#)	gene name (MIM)	location	reported implication
1	C677T (V36,1801133)	MTHFR (607093)	1p36.3	risk factor in vascular disease
1	Arg506Gln (V53, 6025)	F5 (227400)	1q23	activated protein C resistence
1	Ser128Arg (V61, 5361)	SELE (131210)	1q23-25	coronary artery disease
1	Leu554Phe (V62, 5355)	SELE (131210)	1q23-25	coronary artery disease
1	Met235Thr (V42, 699)	AGT (106150)	1q42-43	hypertension
1	Val7Met (V43, 664)	NPPA (108780)	1p36.2	hypertension
1	T2238C (V44, 2238)	NPPA (108780)	1p36.2	
2	Thr71Ile (V8, 1367117)	APOB (107730)	2p24	increased plasma LDL cholesterol level
2	Arg3500Gln (V9, 5742904)	APOB (107730)	2p24	hypercholesterolemia
3	Pro12Ala (V19, 1801282)	PPARG (601487)	3p25	nonsynonymous change, type 2 diabetes
3	A1166C (V41, 5186)	AGTR1 (106165)	3q21-25	hypertension
4	Gly460Trp (V45, 4961)	ADD1(102680)	4p16.3	hypertension
4*	G-455A (V58, 1800790)	FGB (134830)	4q28	progression of atherosclerosis
5	Arg16Gly (V49, 1042713)	ADRB2 (109690)	5q32-34	asthma
5	Gln27Glu (V50, 1042714)	ADRB2 (109690)	5q32-34	obesity
5*	G873A (V59, 1062535)	ITGA2 (192974)	5q23-31	GP Ia/IIa surface expression
6	C93T (V4, 1652503)	LPA (152200)	6q27	atherosclerosis
6	Thr26Asn (V32, 1041981)	LTA (153440)	6p21.3	myocardial infarction
6	Thr26Asn (V68, 1041981)	TNFb (153440)	6p21.3	myocardial infarction
6	G-376A (V64, 1800750)	TNF (191160)	6p21.3	malaria
6	G-308A (V65, 1800629)	TNF (191160)	6p21.3	asthma
6*	G-244A (V66, 673)	TNF (191160)	6p21.3	
6*	G-238A (V67, 361525)	TNF (191160)	6p21.3	protective against type 2 diabetes
6*	G121A (V5, 1800769)	LPA (152200)	6q27	
7	A-922G (V37, 1800779)	NOS3 (163729)	7q36	
7	C-690T (V38, 3918226)	NOS3 (163729)	7q36	
7	Glu298Asp (V39, 1799983)	NOS3 (163729)	7q36	hypertension, Alzheimer's disease
7	5G(-675)4G (V56, 1799768)	PAI1 (173360)	7q21.3-22	coronary artery disease
7	Met55Leu (V25, 3202100)	PON1 (168820)	7q21.3	cardiovascular disease
7	Gln192Arg (V26, 662)	PON1 (168820)	7q21.3	coronary artery disease
7	Ser311Cys (V27, 7493)	PON2 (602447)	7q21.3	coronary artery disease
7*	G11053T (V57, 7242)	PAI1 (173360)	7q21.3-22	

Table 1: (Continued)

block	variant (symbol, dbSNPrs#)	gene name (MIM)	location	reported implication
8	Trp64Arg (V18, 4994)	ADRB3 (109691)	8p12-11.2	NIDDM in some population
8	T-93G (V21, 1800590)	LPL (238600)	8p22	combined hyperlipidemia
8	Asp9Asn (V22, 1801177)	LPL (238600)	8p22	combined hyperlipidemia
8	Asn291Ser (V23, 268)	LPL (238600)	8p22	combined hyperlipidemia
8	Ser447term (V24, 328)	LPL (238600)	8p22	type 1 hyperlipidemia
9	Thr347Ser (V6, 675)	APOA4 (107690)	11q23	
9	Gln360His (V7, 5110)	APOA4 (107690)	11q23	the metabolism of APOB
9	C-641A (V10, 2542052)	APOC3 (107720)	11q23	
9	C-482T (V11, 2854117)	APOC3 (107720)	11q23	increased plasma triglyceride levels
9	T-455C (V12, 2854116)	APOC3 (107720)	11q23	increased plasma triglyceride levels
9	C1100T (V13, 4520)	APOC3 (107720)	11q23	increased plasma triglyceride levels
9	C3175G (V14, 5128)	APOC3 (107720)	11q23	increased plasma triglyceride levels
9	T3206G (V15, 4225)	APOC3 (107720)	11q23	
9*	5A(-1171) 6A (V51, 3025058)	MMP3 (185250)	11q23	coronary heart disease
9*	G20210A (V52, 1799963)	F2 (176930)	11p11-q12	hyperprothrombinemia
10	Trp493Arg (V46, 5742912)	SCNN1A (600228)	12p13	nonsynonymous change
10	Thr663Ala (V47, 2228576)	SCNN1A (600228)	12p13	nonsynonymous change
10	C825T (V48, 5443)	GNB3 (139130)	12p13	hypertension
11	-323 10-bp Ins/Del (V54, 5742910)	F7 (227500)	13q34	hypertension
11	Arg353Gln (V55, 6046)	F7 (227500)	13q34	myocardial infarction
12	C-480T (V20, 1800588)	LIPC (151670)	15q21-23	regulation of plasma lipids
13	C-631A (V29, 1800776)	CETP (118470)	16q21	
13	Ile405Val (V31, 5882)	CETP (118470)	16q21	plasma HDL cholesterol level
13	Asp442Gly (V33, 2303790)	CETP (118470)	16q21	CETP deficiency
13	G+1A (V34, 5742907)	CETP (118470)	16q21	CETP deficiency
13*	C-629A (V30, 1800775)	CETP (118470)	16q21	
14	Alu-element Ins/Del (V40, 1799752)	ACE(or DCP1) (106180)	17q23	myocardial infarction
14	Leu33Pro (V60)	ITGB3 (173470)	17q21.32	coronary heart disease
15	Cys112Arg (V16, 429358)	APOE (107741)	19q13.2	hyperlipoproteinemia
15	Arg158Cys (V17, 7412)	APOE (107741)	19q13.2	hyperlipoproteinemia
15	Gly241Arg (V63, 1799969)	ICAM1 (147840)	19p13.3-13.21	
15	NcoI+/- (V28, 5742911)	LDLR (606945)	19p13.2	cholesterol homeostasis
16	Ile278Thr (V35, 5742905)	CBS (236200)	21q22.3	homocystinaria

* means that the locus is not used in data analysis due to the problem of heavy missing in the sense that at least 21% of the 136 individual genotypes are imcomplete at this locus.

Table 2: Haplotype-interactions that predispose to type 1 diabetes

Block-pair	Chromosomal location	Cases			Control		
(variant-pair)	(gene-pair)	p-value	q-value	z-score	p-value	q-value	z-score
(1,4)	(1,4)	0.000	0.000	-2.9732	0.40	0.13	-0.133
(V44:V45)	(NPPA:ADD1)	0.01		-3.4988	0.964		1.0808
(1,10)	(1,12)	0.000	0.000	-3.5908	0.22	0.1	-0.712
(V62:V46)	(SELE:SCNN1A)	0.032		-5.8305	1.00		0.3251
(1,15)	(1,19)	0.000	0.000	-3.7758	0.60	0.18	0.243
(V43:V17)	(NPPA:APOE)	0.018		-3.2417	0.93		0.3975
(2,7)	(2,7)	0.31	0.12	-0.4364	0.01	0.019	-2.4063
(V8:V39)	(APOB:NOS3)	0.35		-0.2049	0.002		-3.7312
(2,8)	(2,8)	0.03	0.036	-2.0210	0.20	0.10	-0.709
(V9:V22)	(APOB:LPL)	0.002		2.3361	1.00		-0.9950
(2,12)	(2,15)	0.032	0.037	-2.2290	0.59	0.18	0.315
(V8:V20)	(APOB:LIPC)	0.6		0.4021	0.022		-2.2037
(3,15)	(3,19)	0.258	0.11	-0.6140	0.00	0.000	-4.380
(V41:V16)	(AGTR1:APOE)	0.528		0.1457	0.0475		-2.1793
(5,7)	(5,7)	0.22	0.10	-0.9170	0.000	0.000	-2.2606
(V50:V38)	(ADRB2:NOS3)	0.37		-0.3814	0.01		-2.7346
(6,7)	(6,7)	0.01	0.017	-2.4328	0.30	0.12	-0.6584
(V4:V26)	(LPA:PON1)	0.0075		-3.1221	0.996		1.3345
(V4:V37)	(LPA:NOS3)	0.016		-3.0343	0.968		0.7989
(V4:V39)	(LPA:NOS3)	0.636		0.5031	0.046		-2.6121
(V65:V38)	(TNF:NOS3)	0.014		-2.5465	0.85		1.0279
(V68:V37)	(TNFb:NOS3)	0.01		-3.6194	0.222		-0.6165
(6,16)	(6,21)	0.45	0.15	-0.4934	0.000	0.000	-2.0156
(V32:V35)	(TNFb:CBS)	0.61		-0.2664	0.0375		-4.2425
(7,9)	(7,11)	0.21	0.10	-0.9432	0.000	0.000	-2.9033
(V38:V7)	(NOS3:APOA4)	0.16		-1.6673	0.000		-11.3963
(V26:V7)	(PON1:APOA4)	0.37		-0.3737	0.025		-2.3077
(V25:V10)	(PON1:APOC3)	0.51		-0.0683	0.028		-2.4287
(V25:V12)	(PON1:APOC3)	0.78		0.8031	0.028		-2.1233
(V38:V11)	(NOS3:APOC3)	0.786		0.6522	0.004		-3.5639
(V37:V11)	(NOS3:APOC3)	0.78		0.6673	0.028		-2.6518
(V37:V10)	(NOS3:APOC3)	0.296		-0.4199	0.002		-4.8354
(V37:V12)	(NOS3:APOC3)	0.73		0.7839	0.000		-4.7122
(V38:V13)	(NOS3:APOC3)	0.71		0.4653	0.024		-3.3368
(9,13)	(11,16)	0.02	0.027	-2.3164	0.19	0.1	-0.7887
(10,12)	(12,15)	0.155	0.09	-0.9611	0.01	0.019	-2.5475
(V47:V20)	(SCNN1A:LIPC)	0.172		-1.1070	0.002		-4.7115
(14,16)	(17,21)	0.000	0.000	-4.1709	0.59	0.18	0.1466
(V40:V45)	(DCP1:CBS)	0.025		-4.0039	0.697		0.7439

Table 3: P-values and z-scores for testing interactions within Block 1

Splitting-location	Cases		Control	
	p-value	z-score	z-value	z-score
1	0.42	-0.0497	0.23	-0.6620
2	0.06	-1.5060	0.40	-0.3271
3	0.07	-1.6728	0.69	0.5059
4	0.00	-2.4910	0.31	-0.4083
5	0.05	-1.6461	0.53	0.0784
6	0.10	-1.2424	0.33	-0.2145

Table 4: Within-haplotype-interactions that predispose to type 1 diabetes

Block	splitting-location	Cases		Control	
		p-value	z-score	p-value	z-score
1	4	0.00	-2.4910	0.31	-0.4083
	5	0.05	-1.6461	0.53	0.0784
6	3	0.27	-0.5888	0.00	-4.2116
7	5	0.04	-1.9404	0.26	-0.5867
8	1	0.16	-0.7903	0.01	-2.9411
	2	0.52	0.2147	0.02	-3.0241
9	3	0.11	-1.3035	0.00	-7.8697
	4	0.10	-1.1386	0.00	-9.2839
	5	0.13	-0.8928	0.00	-3.6252
	6	0.19	-0.8848	0.00	-4.8922
	7	0.46	-0.1373	0.02	-1.7784

Table 5: The result of the stepwise logistic-regression for the Dutch data

Terms added sequentially				
Variant	Coefficient	Std. error	Resid. deviance	p-value
Intercept	0.93277	0.42811	175.35	
V41	-0.04511	1.10738	170.65	0.0302
V4	17.2308	1.10738	162.98	0.0056
V26	-1.3954	0.94438	158.65	0.0375
V64	0.73047	1.21786	151.01	0.0057
V15	4.58316	1.49003	141.54	0.0021
V9(missing)	19.6659	73.3194	137.1299	0.0357
V28	-4.8015	1.9924	129.9424	0.0073
V41:V64	-11.0055	4.3711	122.5526	0.0066

Figure legends

Figure 1 The p-values and z-scores for 40 sets of genotypes with $(\theta, R) = (4, 0)$, $(4, 4)$, $(4, 20)$ and $(16, 16)$ respectively. The dotted lines are for the data sets where there are interactions between two haplotype-blocks, while the lines with small triangles are for the data sets where there are no interactions between the two haplotype-blocks.

Figure 2 The p-values of testing the interactions of Blocks 1$\sim$8 with the other blocks for the cases and controls in the Dutch type 1 diabetes data, respectively. The dotted lines are for the cases while the lines with small triangles are for the controls. The normal line, a contrast between the two lines is derived by subtracting the corresponding p-values of the control from those of the case.

Figure 3 The p-values of testing the interactions of Blocks 9$\sim$16 with the other blocks for the cases and controls in the Dutch type 1 diabetes data, respectively. The dotted lines are for the cases while the lines with small triangles are for the controls. The normal line, a contrast between the two lines is derived by subtracting the corresponding p-values of the control from those of the case.

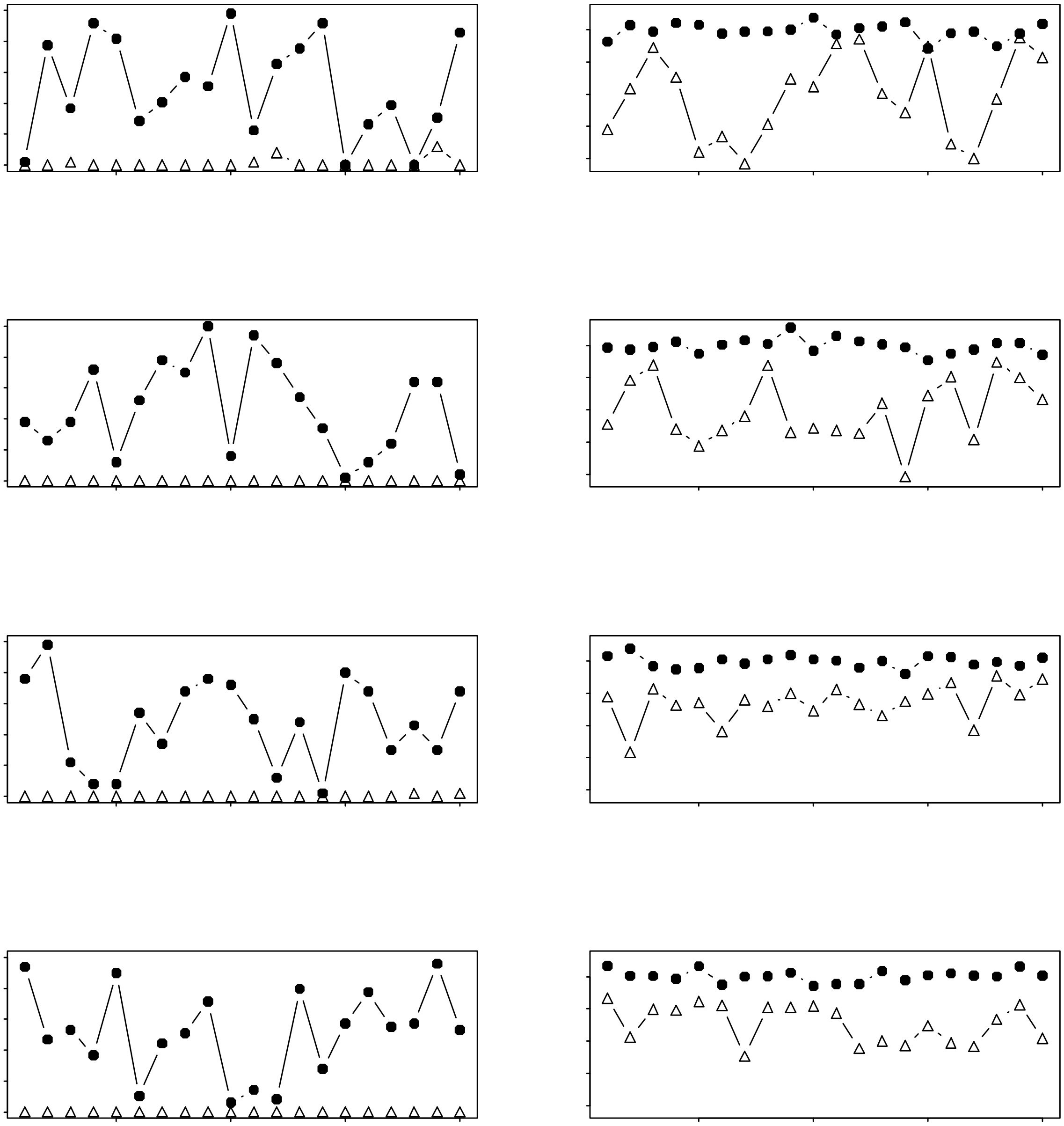

Fig. 1.—

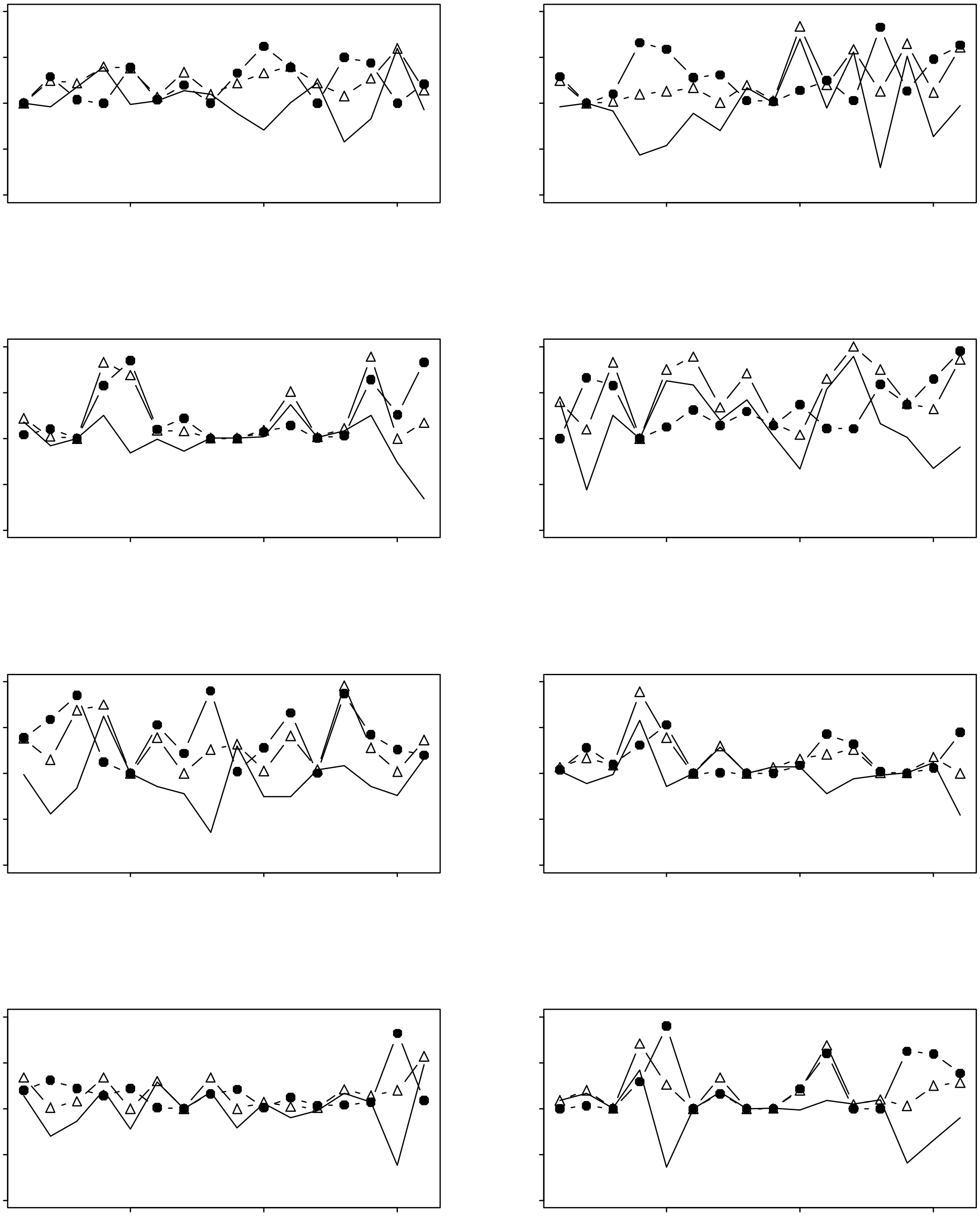

Fig. 2.—

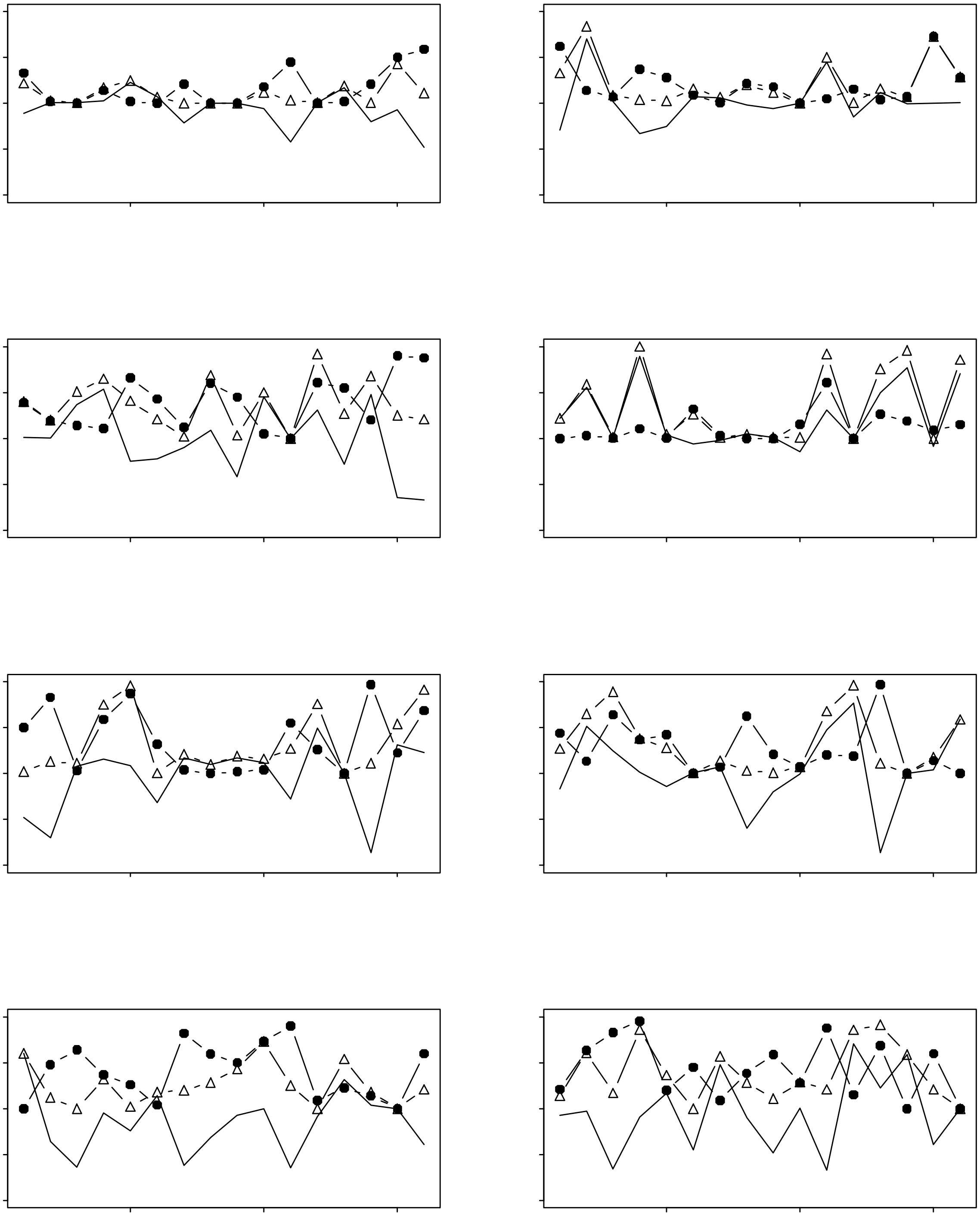

Fig. 3.—