Fine genetic mapping using haplotype analysis and the missing data problem

M. N. CHIANO[1,2*] AND D. G. CLAYTON[1]
[1]*MRC – Biostatistics Unit, Institute of Public Health, Cambridge*
[2]*CRC Human Cancer Genetics Research group, University of Cambridge*

(Received 13.10.97. Accepted 26.1.98)

SUMMARY

The genetic basis of many human diseases, especially those with substantial genetic determinants, has been identified. Notable amongst others are cystic fibrosis, Huntington's disease and some forms of cancer. However, the detection of genetic factors with more modest effects such as in bipolar disorders and a majority of the cancers, has been more complicated. Standard linkage analysis procedures may not only have little power to detect such genes but they do, at best, only narrow the location of the disease susceptibility gene to a rather large region. Association studies are therefore necessary to further unveil the aetiological relevance of these factors to disease. However, the number of tests required if such procedures were used in extended genome-wide screens, is prohibitive and as such association studies have seen limited application, except in the investigation of candidate genes. In this paper, we discuss a logistic regression approach as a generalization of this procedure so that it can accommodate clusters of linked markers or candidate genes. Furthermore, we introduce an expectation maximization (E–M) algorithm with which to estimate haplotype frequencies for multiple locus systems with incomplete information on phase.

INTRODUCTION

When a mutation is first introduced into a population, it resides on a single chromosome and, therefore, on a single background haplotype of some genetic landmarks (markers). At this stage, there is complete linkage disequilibrium between these markers and the disease mutation, i.e. a non-random association of alleles at the linked loci. Consequently, one finds the disease mutation *only* in the presence of a specific set of marker alleles. When the loci are far apart, this association is easily eroded over the generations due to recombination. However for tightly linked loci, this relationship will be conserved, at least, over a considerable number of generations.

In the study of simple Mendelian diseases, in particular, rare traits for which it is difficult to assemble a corroborative set of recombination events, haplotype analysis has been a great asset in localizing these genes. For instance, tracing the cosegregation of a disease and the marker haplotypes in families that independently support linkage can reveal key recombinants which may exclude those regions of the genome deemed to be incompatible with the genetic model and may suggest flanking markers to the disease locus, (see for example Goudie *et al.* 1993; Nellist *et al.* 1993 and Povey *et al.* 1994).

Many common diseases, however, are genetically heterogeneous with the same clinical manifestation caused by different susceptibility genes or a combination of low penetrant genes with possible environmental effects. Some diseases, such as Multiple Sclerosis, have late age at onset and others may directly or indirectly interfere with reproductivity in which case, clusters of high risk families are difficult to find. Although sib-pair analysis methods offer a powerful alternative statistical tool in mapping such diseases

[*] Present address for correspondence: Dr M. N. Chiano Twin Research Unit, St Thomas' Hospital, Lambeth Palace Road, London SE1 7EH, U.K., Tel.: + 44 171 928 9292 ext. 1573. Fax: + 44 171 922 8154.
E-mail: m.chiano@umds.ac.uk

and those whose segregation model is uncertain, there are many diseases for which it is difficult to ascertain a reasonable number of sibpairs.

The availability of Microsatellite DNA polymorphisms has made it possible to identify markers in almost any region of the genome and the human map is growing ever denser. This development has led to many studies that sought to demonstrate association between disease phenotype and DNA markers at or near genes of interest (candidate gene approach). A positive finding for association is taken as tentative evidence that the marker is linked to the disease susceptibility gene or perhaps, that it is the 'candidate' gene itself. Efficient methods for testing disease association with bi-allelic candidate gene systems are discussed elsewhere (Chiano & Clayton, 1998). With highly polymorphic systems, however, tests for association based on pairwise comparisons are fraught with difficulty. For example, in testing for disease association with the major histocompatibility complex (MHC), there is often no reason, *a priori*, to single out any particular antigen for scrutiny and it would be prudent to test for association of disease with each of the antigens in turn. With 50 or so tests on the same sample, almost any study is likely to uncover at least one positive result, even when none exists. Not only are the tests numerous, but they are also not independent because non-allelic genes are likely to segregate together by virtue of their proximity. Hence, the usual test size adjustment problems.

With the available statistical methods, it is difficult to determine whether a particular allele is directly involved in disease susceptibility (primary factor) or it marks the effect of other linked gene(s) (secondary factor(s)). Within the MHC, for example, many studies have reported clusters of adjacent loci that 'occur' together as sets, in which case, it is not possible to isolate the effect of any particular locus or allele from the rest (Dawkins *et al.* 1983; Bodmer *et al.* 1986). Moreover, the situation in certain diseases is even more complex – many alleles at multiple loci are found to be associated with disease but with entirely different alleles in different sub groups; some alleles may be susceptibility genes while the rest are suppressors and so on. This type of heterogeneity has been reported in the study of insulin dependent diabetes mellitus (IDDM) (Hawkins *et al.* 1987; Apricio *et al.* 1988). Therefore, an effective approach to the mapping of such genes should take into account the fact that several genes might be involved, i.e. the haplotype in its entirety should be mapped.

Here, we introduce a general modelling approach which not only tests for association but allows for the correlation within the whole genetic system. This approach can be regarded as exploratory and thus attempts to determine how far the haplotype harbouring the putative disease gene extends. This modelling approach is flexible and is easily extendable to assess the possibility of gene/gene and gene/environment interactions. With Microsatellite and repeat sequence polymorphisms, the data are likely to be sparse and ordinary logistic regression approaches immediately run into difficulty. To accommodate these class of problems, this logistic regression approach is easily recast into a generalized linear mixed model framework with random effects and the parameters of interest estimated by Gibbs sampling. An expectation maximization (E–M) algorithm to obtain haplotype frequencies for multiple locus systems with incomplete information on phase is also discussed. This is further generalized to cope with genetic systems with missing phenotype information at one or more loci. Practical and worked examples applying the procedures and methods discussed here will be presented as a separate paper.

MODEL AND METHOD

(*a*) *Phase known*

Consider a simple three locus system denoted as X, Y and Z. For now, it may be assumed that there is complete data, i.e. that the parental genotypes of all subjects in the study are known and we can infer the possible haplotypes accurately. Let the l_j alleles at each locus be the l-levels of the jth random variable or covariate (here, $j = 1, ..., 3$). For simplicity, let us assume a simple qualitative trait phenotype. In other words, we suppose that each sampled individual is classified either as affected (case) or unaffected (control), i.e.

$$D = \begin{cases} 0 & \text{Unaffected} \\ 1 & \text{Affected,} \end{cases}$$

and, consequently, haplotypes are classified as to whether they are derived from a case or a control. Furthermore, we assume that the cases and controls are randomly sampled from the study population. It should be noted however, that with properly designed matched case/control studies, the methods discussed here can easily extend to transmission/disequilibrium testing by fitting conditional logistic regression models (Self *et al.* 1991) rather than ordinary logistic regression models.

Using the notation above, we have $\prod_{j=1}^{3} l_j$ possible haplotypes with the disease status taken as the response variable. The observed genotypic information at each locus is therefore treated as the dose or the 'stimulus' factor. Let π_{xyz} be the probability vector that haplotype H_{xyz} carries the disease gene, then

$$\pi_{xyz} = P(D = 1|H_{xyz}),$$

and

$$1 - \pi_{xyz} = P(D = 0|H_{xyz}),$$

where the subscripts designate the alleles at each corresponding locus, i.e. $x = 1, ...l_1$; $y = 1, ...l_2$ and $z = 1, ...l_3$. The odds, λ_{xyz}, is then defined as

$$\lambda_{xyz} = \frac{\pi_{xyz}}{1 - \pi_{xyz}}.$$

If the risk attributable to each haplotype is a multiplicative function of its genotypic constituents, we may want to examine the main effects of each locus and the risk associated with each allele by fitting a logistic regression model to the data. Assuming that alleles at different loci are functionally unrelated, the log odds is then represented as

$$\eta_{xyz} = \log(\lambda_{xyz}) = \mu + \alpha_x + \beta_y + \gamma_z, \qquad (1)$$

the main effects model, with

$$\pi_{xyz} = \frac{\exp(\mu + \alpha_x + \beta_y + \gamma_z)}{1 + \exp(\mu + \alpha_x + \beta_y + \gamma_z)}. \qquad (2)$$

If m_{xyz} haplotypes are observed of which n_{xyz} are derived from case chromosomes, the log likelihood is

$$l(\mathbf{\Phi}; X, Y, Z) = \sum n_{xyz}\eta_{xyz} - \sum m_{xyz} \log(1 + \exp(\eta_{xyz})), \qquad (3)$$

where $\mathbf{\Phi} = (\alpha, \beta, \gamma)$ are the main effects of locus X, Y, Z (McCullagh & Nelder, 1989). In principle, this approach is flexible and one can easily assess the relative contribution to disease aetiology of the individual polymorphisms while allowing for the possible effects of other loci. For complex diseases, however, functional independence between alleles at different loci is likely to be an exception rather than the rule and testing for disease association with each variant in turn may yield less impressive results. In this instance, increase in disease susceptibility may be associated with a combination of alleles at one or more of these loci. These associated alleles constitute the *haplotypes* and their effects are assessed by fitting interaction terms in the logistic regression model. The level of interaction therefore suggests how far the haplotype harbouring the putative disease gene(s) extends.

(*b*) *Phase unknown*

In the study of diseases where parents are easy to find, e.g. IDDM, the genotypic constitution of each case can be inferred with reasonable accuracy from the parental marker phenotypes, and the number of case haplotypes determined simply by counting. However, in late onset traits (e.g. most cancers) and those that affect reproductivity (for instance, multiple sclerosis (MS)), the above approach is not always possible. Determining case and control haplotypes in the absence of phase information poses an estimation problem but is easily solved using standard missing data techniques such as the E–M algorithm, briefly discussed below.

In general, suppose we have N population cases and a comparable number of controls, say N'. Let each subject be typed for a set of polymorphisms, K, each of which has l alleles ($l \geqslant 2$ and not necessarily all equal). If the phase for each subject was known (the true genotypic constitution), its two haplotypes would be uniquely determined. However without phase, the complementary pair of haplotypes from each subject is a realisation from a set of 2^{K-1} possible constituent genotypes.

As before, suppose that each case, i, is typed for three relatively close polymorphisms with the following phenotypic constitution $(a, b), (c, d)$ and (e, f) at the three loci, respectively. The 'genotypic constitution' of each individual at a particular locus is defined as the locus genotype including phase information. Without phase, therefore, any subject (i) with the above multi-locus phenotype would have any of four possible genotypic constitutions, namely:

Gen	$H^{(i)}_{xyz}$	$\tilde{H}^{(i)}_{xyz}$
1	$a - c - e$	$b - d - f$
2	$a - d - e$	$b - c - f$
3	$a - c - f$	$b - d - e$
4	$a - d - f$	$b - c - e$

where $H^{(i)}_{xyz}$ and $\tilde{H}^{(i)}_{xyz}$ is the pair of complementary haplotypes corresponding to each genotypic constitution with alleles x, y, z $(x, y, z = 1, 2...l)$ at the three loci, respectively. Denote the corresponding haplotype frequencies as $\theta^{(i)}_{xyz}$ and $\tilde{\theta}^{(i)}_{xyz}$. In addition, let the complete listing of all haplotypes compatible with the multi-locus phenotype for case i be denoted as $\mathscr{E}^{(i)}_{\text{Gen}}$, i.e. the set of genotypic constituents of i. Henceforth, parameters or sets of parameters without the superscript will refer to the entire sample rather than a single subject. For example, $\mathscr{E}_{\text{Gen}}$ is the set of all possible genotypes in the sample while $\mathscr{E}^{(i)}_{\text{Gen}}$ is the set of genotypes compatible with the phenotypic constitution of subject i.

First, we assume that each complementary pair of haplotypes is equally likely. This is analogous to setting the initial phase parameter to one half. Therefore, the probability that subject i with phenotypic constitution $w^{(i)}$ would have the above genotypic make-up is given by

$$P(w^{(i)}) = \sum_{\mathscr{E}^{(i)}_{\text{Gen}}} \theta^{(i)}_{xyz}.\tilde{\theta}^{(i)}_{xyz} \tag{4}$$

and the log likelihood function for the sample is

$$l(w) = \sum_{i=1}^{N} \log \left\{P(w^{(i)})\right\}. \tag{5}$$

If $n^{(i)}_{xyz}$ is the contribution of subject i to haplotype H_{xyz}, then

$$E(n^{(i)}_{xyz}) = \frac{\theta^{(i)}_{xyz} \cdot \tilde{\theta}^{(i)}_{xyz}}{P(w^{(i)})} \tag{6}$$

and the total contribution from the entire sample is

$$n_{xyz} = \sum_{i=1}^{N} E(n^{(i)}_{xyz}), \quad \text{with} \quad \sum n_{xyz} = 2N. \tag{7}$$

Applying the same procedure to the N' controls yields corresponding quantities (n'_{xyz}) in the control set. In E–M jargon, this is the expectation step (E-step).

The log likelihood $l(\boldsymbol{\Theta}; H_{xyz})$, is then maximized and the initial haplotype frequencies are updated to

$$\theta^{*}_{xyz} = \frac{n^{*}_{xyz}}{\sum n^{*}_{xyz}} = \frac{n^{*}_{xyz}}{2N}, \tag{8}$$

where n^{*}_{xyz} are fitted values from the logistic regression model. The above procedures are then repeated until the parameters stabilize. In practice, the best linear predictor (η_{xyz}), is not known. A way round this problem is to fit a 'saturated' model to the data replacing the M-step with a simple counting procedure. Maximum likelihood estimates of the number of case and control haplotypes are then used to search for a parsimonious linear logistic model from which refined parameter estimates are obtained.

The algorithm as described pre-supposes that each subject has phenotype information at all 3 loci. This is hardly the case in practice but this is easily extendable to cope with cases that have partial or missing phenotype information. For example, a subject with missing information at locus j is a realization from a set of $l_j(l_j + 1)/2$ possible phenotypic constitutions, where l_j is the number of alleles at the the missing locus. In general, a subject with missing information at L loci is a realization from M admissible phenotypic constitutions, where $M = \prod_{j=1}^{L} l_j(l_j + 1)/2$. Denote this set as $\mathscr{E}_M^{(i)}$. The E–M algorithm is modified to handle such cases by looping over this set of compatible phenotypic constitutions for each subject with missing information. Hence, equation (4) becomes

$$P(w^{(i)}) = \sum_{\mathscr{E}_M^{(i)}} \sum_{\mathscr{E}_{\text{Gen}}^{(i)}} \theta_{xyz}^{(i)} . \tilde{\theta}_{xyz}^{(i)} \tag{9}$$

and the iteration proceeds as before. However, this modification requires substantially more computer time and memory. These algorithms are available in S code.

The primary difficulty in applying this regression approach is the large number of potential haplotypes in highly polymorphic systems, in which case, fixed effect models break down due to the large number of effects that have to be estimated. However, with the availability of cheap computer power, we can recast the regression model (1) as a generalized linear mixed model (GLMM) with random effects. The full model, including possible environmental covariates, has many parameters and may be intractable. However, this model is easily fitted using Gibbs Sampling which reduces to sampling, from full conditional distributions, each parameter given the data and current estimates of all other model parameters (see Clayton, 1995; Thomas *et al.* 1995). In a recent study of the association between the HLA system and IDDM, Thomas *et al.* (1995) employ this procedure, but instead they consider only the 'saturated' term for haplotype effects. Although this model might be adequate when the extent of the haplotype is known, we envisage that such an approach may yield spurious results especially when the entire haplotype does not necessarily make a genetic contribution to disease. An elaborate hierarchical model with random effects of haplotypes such as the following, is much more appropriate

$$\log\,(\lambda_{xyz,E_e}) = \mu + \underline{\beta}_G + \underline{\beta}_{E_e} + \underline{\beta}_{G*E_e} + \underline{\beta}_H + \underline{\beta}_{H*E_e}, \tag{10}$$

where G and E_e, the genotype and environmental covariates, define the fixed-effect part of the model and $\underline{\beta}_w \sim \mathscr{N}(\mathbf{0}, \mathbf{\Lambda}_w)$ (w stands for $H, H*E_e$, the haplotype and possible haplotype-environment interaction effects) constitute the random-effect part; $\mathbf{\Lambda}_w$ being the precision matrix for the corresponding model parameters (see Clayton, 1995). Estimates of $\mathbf{\Lambda}_H$ elucidate the extent of the haplotype while Bayes' estimates of random effects identify the high risk haplotypes. In the case of HLA and other highly polymorphic systems, there is usually a large number of possible haplotypes some of which may be rare and or correlated. When this correlation is known, *a priori*, this can be accounted for by considering more structured prior precision matrices, $\mathbf{\Lambda}_w$ (some of these issues are the subject of a separate investigation).

DISCUSSION

We have considered the linear logistic regression model as an appropriate tool for locating susceptibility genes, and investigating possible gene/gene and gene/environment interaction effects, simultaneously. Also, we have introduced an E–M type algorithm for estimating haplotype frequencies in population-based samples with missing information on phase. This algorithm is fairly general and will utilize information from individuals with missing phenotype at any of the loci being investigated. The regression approach is flexible and can easily be extended to accommodate elaborate and more complicated models including environmental covariates using Gibbs Sampling.

This work was funded by a programme grant from the Cancer Research Campaign (CRC).

REFERENCES

Aparicio, J., Wakisaka, A., Takada, A., Matsura, N. & Aizawa, M. (1988). HLA–DQ system and insulin-dependent diabetes mellitus in Japanese: does it contribute to the development of diabetes as it does in Caucasians. *Immunogenetics* **28**, 240–246.

Bodmer, W., Trowsdale, J., Young, J. & Bodmer, J. (1986). Gene Clusters and the evolution of the major histocompatibility system. *Phil. Trans. R. Soc. of Lond.* [*biol.*] **312**, 303–315.

Chiano, M. N. & Clayton, D. G. (1998). Genotypic relative risks under ordered restriction. *Genetic Epidemiol.* (in press).

Clayton, D. G. (1995). Generalized linear mixed models. In *Markov Chain Monte Carlo in Practice* (ed. W. R. Gilks, S. Richardson & D. J. Spiegelhalter), pp. 275–301. London: Chapman & Hall.

Dawkins, R., Christiansen, F., Kay, P., Garlepp, M., McCluskey, J., Hollingsworth, P. & Zilko, P. (1983). Disease associations with complotypes, supratypes and haplotypes. *Immunol. Rev.* **70**, 5–22.

Goudie, D., Yuille, M., Laversha, M., Furlong, R., Carter, N., Lish, M., Affara, N. & Ferguson-Smith, M. A. (1993). Multiple self-healing squamous epithelioma (ESS1) mapped to chromosome 9q22–q31 in families with a common ancestry. *Nature Genetics* **3**, 165–169.

Hawkins, B., Lam, K., Ma, J., Low, L., Cheung, P., Serjeanston, S. & Yeung, R. (1987). Strong association of HLA-DR3/DR9 heterozygosity with early-onset insulin-dependent diabetes mellitus in Chinese. *Diabetes* **36**, 1297–1300.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models* (2nd edition). London: Chapman & Hall.

Nellist, M., Brook-Carter, P., Connor, J., Kwiatkowski, D. & Johnson, P. (1993). Identification of markers flanking the tuberous sclerosis locus on chromosome 9 (TSC1). *J. Med. Genet.* **30**, 224–227.

Povey, S., Burley, W., Attwood, J., Benham, F., Hunt, D., Jeremiah, J., Franklin, D. *et al.* (1994). Two loci for tuberous sclerosis: one on 9q34 and one on 16p13. *Ann. Hum. Genet.* **58**, 107–127.

Self, S. G., Longton, G., Kopecky, K. J. & Liang, K. (1991). On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* **47**, 53–61.

Thomas, D., Janne Pitkaniemi, Langholz, R., Tuomilehto-Wolf, E., Tuomilehto, J. and the DiMe Study Group (1995). Variation in HLA-associated risks of childhood insulin-dependent diabetes in the Finnish population. II. Haplotype effects. *Genetic Epidemiol.* **12**, 455–466.