SHORT REPORT

The use of Mx for association and linkage analysis

Michael C. Neale

Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Box 980126, Richmond, VA 23298-0126 U.S.A.

Abstract

Evidence for genetic linkage, obtained from a correlation between phenotypic similarity and genetic similarity at a specific chromosomal location typically yields a broad genomic region in which a candidate locus might be found. Evidence for association is usually gathered from case control studies and is subject to false positives from phenomena such as population stratification. Data from relatives may be used to distinguish population stratification from genuine allelic effects in an association context. Of special interest is joint linkage and association which may be used for fine mapping because evidence for linkage will be reduced in the presence of evidence for association. This article describes the implementation of these methods using Mx, in both path diagram and script formats, and discusses a number of possible extensions to the model.

Keywords association, disequilibrium, fine mapping, linkage, multivariate, stratification.

Received 23 August 2000; accepted XX XXXX 2000

1

Introduction

Initial evidence for the effects of genetic factors on a quantitative or qualitative trait usually comes from a genetically informative epidemiological study of twins or adoptees.[1,2] Subsequently, identification of the specific genetic factors involved may be attempted using one or both of two main strategies. First, linkage studies may reveal positional candidates, by establishing evidence for linkage in specific regions of the genome. Such findings may be compared to information on known genes in the region (available from the human genome project) to establish specific candidate loci. Second, functional candidates may be identified through knowledge of the biochemical or neurological systems involved in the phenotype under study. Functional candidates are often tested using association studies, but they are subject to false positive findings caused by population stratification or insufficient correction for multiple testing.[3]

Evidence for linkage is unlikely to arise by chance, because of the high significance levels typically used in this type of study. Essentially the method works by correlating the degree to which relatives share alleles identical-by-descent (IBD) at an hypothesized position on a chromosome with their degree of phenotypic similarity. Approaches such as Haseman–Elston regression[4] have led to covariance structure models[5–8,9] which are more flexible for multivariate analysis.

However, relatively modest amounts of genetic heterogeneity can cause enormous variability in the location of the peak[10]. Even without heterogeneity the peak lod score is subject to sufficient stochastic variation that 95% confidence intervals on the peak location could encompass a region with many genes. Therefore, methods for fine mapping are highly desirable.

Joint linkage and association analysis[11–13] appears to offer several advantages. This method is really a marriage of the use of identity-by-descent information gathered from examination of a set of linked loci with mean differences associated with specific alleles. It has been shown[11] that the linkage signal decreases while the association signal increases when the alleles being tested are in linkage disequilibrium with (or are at) the trait locus. In this article I describe how Mx[14] may be used to fit this combined linkage and association model, by using either the graphical path diagram representation or the script language. A valuable benefit of this implementation is that it is readily extended to the multivariate and multilocus cases.

Method

Mx statistical model

Mx has a general interface for the analysis of raw ordinal or continuous data by normal theory maximum likelihood.

Correspondence to: Michael C. Neale, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Box 980126, Richmond, VA 23298-0126, U.S.A.
Email: neale@hsc.vcu.edu

Table 1 Expected sib-pair means and differences at a single additive two-allele locus

Genotype		Individual mean		Pair statistics		Stratified means	
Sib 1	Sib 2	Sib 1	Sib 2	Sum/2	Difference/2	Sib 1	Sib 2
A_1A_1	A_1A_1	a	a	a	0	b	b
A_1A_1	A_1A_2	a	0	$a/2$	$a/2$	$b/2 + w/2$	$b/2 - w/2$
A_1A_1	A_2A_2	a	$-a$	0	a	w	$-w$
A_1A_2	A_1A_1	0	a	$a/2$	$-a/2$	$b/2 - w/2$	$b/2 + w/2$
A_1A_2	A_1A_2	0	0	0	0	0	0
A_1A_2	A_2A_2	0	$-a$	$-a/2$	$a/2$	$-b/2 + w/2$	$-b/2 - w/2$
A_2A_2	A_1A_1	$-a$	a	0	$-a$	$-w$	w
A_2A_2	A_1A_2	$-a$	0	$-a/2$	$-a/2$	$-b/2 - w/2$	$-b/2 + w/2$
A_2A_2	A_2A_2	$-a$	$-a$	$-a$	0	$-b$	$-b/w$

Usually, raw numeric data are read using a rectangular Ascii format. Variables may be used to select cases (e.g. select if sex = 1), and a subset of the variables read may be selected for analysis (e.g. **select siblweight sib2weight pihat allelelsibl . . . allele5sib2**). Of the variables selected for analysis, a further sub-setting is possible using the **definition_variables** command. Variables listed in a **definition_variables** command are not selected for analysis, but they may be inserted to modify the elements of any matrix used to define the statistical model. This modification occurs on a case-by-case basis so that the statistical model may be different for every case in the sample. It is this feature that makes it simple to specify random or fixed effects in what is often termed a mixed model.[15,16] Hierarchical linear models[17,18] are also subsumed in this methodology.

In Mx, models may be specified using either the script language or by drawing a path diagram in the graphical user interface which then automatically converts the diagram to a script, executes it, and displays the output parameter estimates and fit statistics in the diagram. Central to both these approaches is a matrix algebra interpreter. It allows the user to declare matrices and to manipulate them using matrix algebra formulae. Predicted means and covariances, case weights, and frequencies may be specified using arbitrarily complex matrix algebra formulae. The normal theory likelihood function L_i of the ith data vector is computed as a frequency-weighted product of a finite mixture[19] of m models:

$$L_i = f_i \prod_{j=1}^{m} w_{ij} g(x_i, \mu_{ij}, \Sigma_{ij})$$

where f_i is the frequency, w_j *is* the weight of model j, and $g(x_i, \mu_{ij}, \Sigma_{ij})$ is the multivariate normal probability density function evaluated at the observed vector x_i for a particular predicted mean vector μ_{ij} and covariance matrix Σ_{ij}. Mx allows the frequencies, the weights, the predicted means and the predicted covariances to be a function of definition variables in each data vector i.

Model for the means

Table 1 shows predicted sibling means and half the pair sum and difference for all possible pairs of siblings classified using a diallelic locus. The homozygote A_1A_1 has a mean of a, the heterozygote has a mean of 0, and the A_2A_2 homozygote has a mean of $-a$. We wish to investigate not only whether the genotypes differ in their means (which is simply a test for the significance of a) but also whether this is as true within sib pairs (who belong to the same stratum) as it is between individuals in different families. Therefore, the model is parameterized separately in terms of the pair means and pair differences, using the symbols a_b for the effect of the A locus between families, and a_w for its effect within sib pairs. By reading the genotype of siblings into Mx and declaring them as definition variables it is possible to select the appropriate coefficients for a_b and a_w from columns five and six of Table 1. The Mx script implements this model using the script language.

Many researchers find path diagrams a useful way to represent models of covariance structure. When drawn correctly, path diagrams provide a mathematically complete description of a model, which makes it possible to specify diagrams using path diagram drawing software such as the Mx graphical user interface (GUI)[14]. Path diagrams consist of circles that represent latent variables, and squares that represent observed variables. These variables are related to each other by causal relations, drawn as single-headed arrows, and by correlational relations drawn as double-headed arrows. It is also possible to construct a model for the means using triangles, which represent constants that do not contribute to the variance of a variable, but only to its mean. One final construct in a path diagram is the placement of an observed, individual-level variable on a path, which is shown as a variable name inside a diamond shape (◇). It is this feature, unique to Mx software at this time, that allows specification of separate models for every sibling pair in the sample. We allow $\hat{\pi}_i$ to differ between pairs in the covariance of the QTL latent variables, and the observed alleles to differ between pairs according to which alleles they have at the locus.

Figure 1 shows a path diagram for a joint linkage and association model. The upper half is the model for the means, which depends on the specific values of the APOE1 and APOE2 definition variables. These variables are the

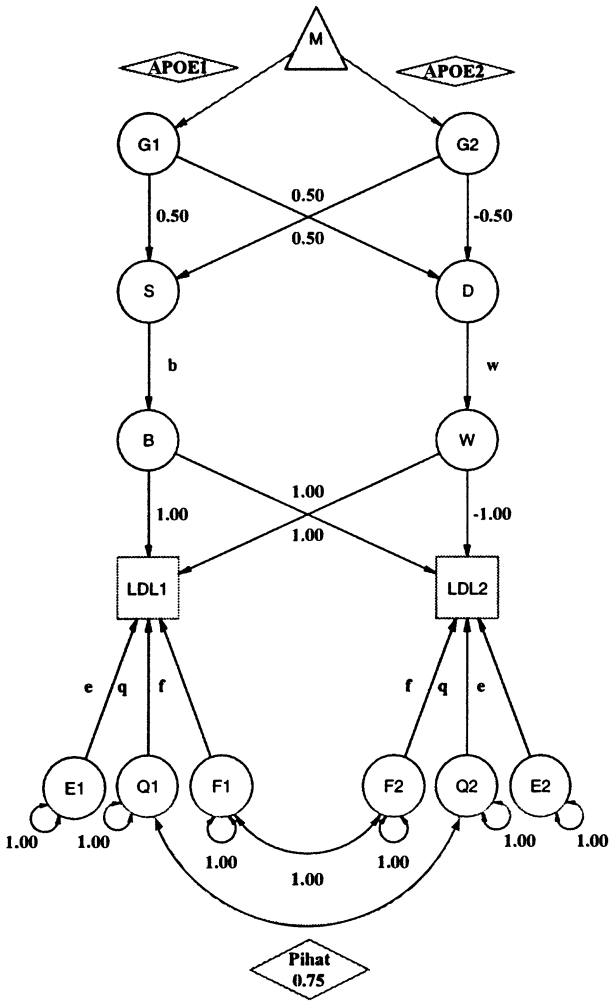

Figure 1 Path diagram of model for joint linkage and association in sib pairs, allowing for possible population stratification. S represents half the sum of the sibling pair's genotypic effects, and D represents its distance. These parameters contribute to between-pair (B) and within-pair (W) effects via parameters *b* and *w*, respectively.

measured genotypes of an individual, scored −1, 0 or 1 according to whether they have zero, one or two alleles of a particular type† at the locus. The mean of an individual is computed by tracing back from their phenotype (LDL1 or LDL2 in Fig. 1) to the constant *M* in the triangle at the top of the page. The values on the paths (1, *b*, *w*, etc.) are multiplied together for each possible pathway from the phenotype to the constant *M*. The predicted mean level is given by the sum of the possible pathways. It is easy to verify that the predicted means of LDLl and LDL2 agree with those in the two right-hand columns of Table 1.

†It is convenient to count the number of 'increasing' alleles, but for multivariate analysis this is not always possible. It is also worth noting that for a locus with many alleles it is possible to compare one allele against all others with no modification to the method.

The lower half of Figure 1 shows the covariance model. This model is limited to three sources of covariation, as would be appropriate for a study of sibling pairs. E_1 and E_2 represent residual variance components specific to each individual, which include environmental factors not shared with a sibling and error of measurement. Q_1 and Q_2 represent the effects of one or more quantitative trait loci that are linked to the location at which the proportion of alleles identical-by-descent (IBD) is computed (usually the locus used in the means model above). The correlation between Q_1 and Q_2 is set to equal $\hat{\pi} = p(IBD = 2) + 0.5p(IBD = 1)$. F_1 and F_2 represent familial factors, both genetic and environmental, shared by siblings, btit which exclude the effects of the locus in the means model and other loci linked to it.

Relevant statistical tests for association are obtained by comparing the log-likelihood of the full model (with parameters *w*, *b*, *f*, *q* and *e* free) against a reduced model. Under certain regularity conditions and the assumption of conditional normality, twice the difference in log-likelihood between the full model and a nested submodel is asymptotically distributed as χ^2 with degrees of freedom (d.f.) equal to the difference in degrees of freedom between the two models. A one-d.f. test for population stratification is given by equating the between- and within-family components *w* and *b*. A conservative test for allelic effects is the one-d.f. test of $w = 0$ against the full model, and a more powerful test may be given by $b = w = 0$ which assumes no stratification. Under some circumstances, this test can be less powerful, if for example stratification exists but it counteracts the effect of the alleles.

Statistical tests for linkage are provided by testing for the significance of the parameter *q* in the model. If there are genuine effects of the locus being used to model the means, eliminating the allelic effects on the means would normally increase the size of parameter *q*, because the linkage signal would include the allele effects. It is this comparison which facilitates fine mapping. The availability of single nucleotide polymorphisms, which are by nature diallelic, makes this approach especially attractive.

Extensions to the method

Perhaps the greatest advantage of the Mx implementation of the Fulker–Cherny joint linkage and association model is that it is very simple to extend to a variety of multivariate cases[20]. The simplest such extension is a common factor model, in which a variety of traits correlate because they share a source of variance in common. Using the method described above, it is straightforward to make the observed variables LDL1 and LDL2 into latent factors, and to specify causal paths to a set of observed measures for each subject. This 'common pathway' or 'psychometric factor' model is a natural extension of the common factor model widely used in psychometrics, econometrics and many other fields. Especially notable in the joint linkage and association model is that the model is specified with a factor mean, which in

turn influences the means of the observed scores. In fact this model makes strong predictions about the degree of influence of the locus on the observed scores, namely that the mean differences between individuals with different genotypes will be proportionate to the size of the loading on the common factor. Such a prediction may or may not be borne out by the data. The association part of the model (in the means) would be accounting for both variation within and covariation between the various traits, and would lead to a reduction in both residual variation and covariation generated by parameters q_i that model the linkage effect on each of the i traits. Multivariate tests of this sort are often more powerful than univariate tests, because covariance as well as variance is explained by the model.

A second straightforward extension is to the multiple locus case, as I described in the volume arising from the first International Meeting on the Genetic Epidemiology of Complex Traits[9]. In practice, linkage studies in humans are unlikely to be of sufficient size to detect epistatic interactions between loci. If they do, it is even more unlikely that it will be possible to discriminate between the different possible types of interaction, namely additive × additive, dominance × dominance and dominance × additive. However, detection of such interactions becomes quite powerful in the context of an association study, where the two interacting loci have been measured. Population stratification notwithstanding, the power for detecting interactions between two diallelic loci should be very similar to a two-way analysis of variance with three levels.

Genetic epidemiology is often much more complicated than the analysis of a single trait together with some genetic marker data or other genetically informative research design such as twins or adoptees. Frequently we seek to understand the action and interaction of risk factors that are associated with outcomes. In many cases we do not have a single indicator of disease status, but must rely on different sources of information about an individual's liability to disease. In child psychiatry it is common to use ratings of children, which may be made by the children themselves, their teachers, parents or other relatives. Some traits show substantial rater bias such that the reports made appear to reflect as much about the individual making the rating as about the one being rated[21]. In such cases it would be especially important to partition variation in the measures into bias and true score components and to focus the genetic model on the true score rather than the bias. Similarly, co-morbidity between disorders may arise for a number of reasons[22] and the choice of an appropriate model could enhance the detection of trait loci. Conversely, correct modelling of the gene action on multiple disorders could enhance the understanding of the sources of co-morbidity.

Gene–environment interactions and correlations provide another area of substantial complexity for genetic epidemiology. G–E interaction occurs when the same genes have different effects on the phenotype depending on the environment in which they are expressed. In a traditional twin study context such effects might be detected by separating each twin group into three subsets (concordant exposed, discordant exposed, concordant non-exposed) according to some dichotomous environmental factor and testing for differences in the heritability between groups. In a linkage/association context the same basic approach could be used to detect interactions at the locus or region in question. It is worthwhile noting that the **definition_variable** technology allows a more flexible analogue of this approach that can handle continuous as well as binary environmental indices.

In summary, one could take any model from a genetic epidemiology textbook such as Neale & Cardon[23] and augment it with both mean allelic and IBD-based QTL effects. These methods form a bridge between three areas: linkage and association studies; traditional genetic epidemiological studies of twins and families; and non-genetic studies of co-morbidity and risk factors. As such, they should prove to be fertile ground indeed for future health research.

References

1 Neale MC. Twin studies. In: Armitage P, Colton T, eds *Encyclopedia of Biostatistics*. New York: John Wiley 1998.
2 Neale MC. Adoption studies. In: Armitage P, Colton T, eds *Encyclopedia of Biostatistics*. New York: John Wiley 1998.
3 Risch N, Zhang H. Mapping quantitative trait loci with extreme discordant sib pairs: Sampling considerations. *Am J Human Genet* 1996; **58**: 836–843.
4 Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a locus. *Behav Genet* 1972; **2**: 3–19.
5 Amos CI, Zhu DK, Boerwinkle E. Assessing genetic linkage and association with robust components of variance approaches. *Ann Hum Genet* 1998; **60**: 143–160.
6 Blangero J, Almasy L. Multipoint oligogenic linkage analysis of quantitative traits. *Genetic Epidemiol* 1997; **14**: 959–964.
7 Eaves W, Neale MC, Maes HH. Multivariate multipoint linkage analysis of quantitative trait loci. *Behav Genet* 1996; **26**: 519–526.
8 Fulker DW, Cherny SS. An improved multipoint sib-pair analysis of quantitative traits. *Behav Genet* 1996; **26**: 527–532.
9 Neale MC. QTL mapping with sib-pairs: the Flexibility of Mx. In: Spector TD, Snieder H, MacGregor AJ, eds *Advances in Twin and Sib-Pair Analysis*. London: Greenwich Medical Media 1999.
10 Roberts S, MacLean C, Neale M *et al*. Replication of linkage studies of complex traits: An examination of variation in location estimates. *Am J Hum Genet* 1999; **65**: 876–884.
11 Cardon LR, Abecasis GR. Some properties of a variance components model for fine-mapping quantitative trait loci. *Behav Genet*., in press.
12 Fulker DW, Cherny SS, Sham PC, Hewitt JK. Combined linkage and association sib pair analysis for quantitative traits. *Am J of Hum Genet* 1999; **64**: 259–267.
13 Neale MC, Cherny SS, Sham P *et al*. Distinguishing population stratification from genuine allelic effects with mx: Association of ADH2 with alcohol consumption. *Behav Genet* 1999; **29**: 233–244.
14 Neale M, Boker S, Xie G, Maes H. *Mx: Statistical Modeling*, 5th edn. Box 980126, Richmond VA. Department of Psychiatry, Virginia Commonwealth University, 1999.
15 Searle S, Casella G, McCulloch C. *Variance Components*. New York John Wiley & Sons, 1992.

16 Vonesh E, Chinchilli V. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker, 1996.
17 Bryk AS, Raudenbush SW. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park: Sage Press, 1992.
18 Prosser R, Rabash J, Goldstein H. *ML3: Software for Three-Level Analysis. User's Guide for v.2*. Institute of Education, University of London, 1991.
19 Everitt BSH, Hand DJ. *Finite Mixture Distributions*. Chapman & Hall, 1981.
20 Eaves LJ, Neale MC, Maes HH. Multivariate multipoint linkage analysis of quantitative trait loci. *Behavior Genetics* 1996; **26**: 519–526.
21 Neale MC, Stevenson J. Rater bias in the EASI temperament scales: A twin study. *J Personality Social Psychol* 1989; 446–455.
22 Neale MC, Kendler KS. Models of comorbidity for multifactorial disorders. *Am J of Hum Genet* 1995; .
23 Neale MC, Cardon LR. *Methodology for Genetic Studies of Twins and Families*. Kluwer Academic Publishers., 1992.

Author Query Form

Journal: Genescreen

Article: 032

Dear Author,

During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.

Many thanks for your assistance.

Query Refs.	Query	Remarks
1	**accepted date not on manuscript.**	
2	**Meaning? 'fits statistics'?**	
3	**Details now available?**	
4	**Place of publication?**	
5	**Vol. no?**	
6	Please supply the vol. no. & page range	
7	**Place of publication?**	