

A logistic mixture model for a family-based association study

Guan Xing¹, Chao Xing^{1,2,3}, Qing Lu¹ and Robert C Elston*¹

Address: ¹Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106, USA, ²Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8591, USA and ³McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8591, USA

Email: Guan Xing - guan.xing@bms.com; Chao Xing - chao.xing@utsouthwestern.edu; Qing Lu - qing.lu@case.edu; Robert C Elston* - rce@darwin.case.edu

* Corresponding author

from Genetic Analysis Workshop 15
St. Pete Beach, Florida, USA. 11–15 November 2006

Published: 18 December 2007

BMC Proceedings 2007, 1(Suppl 1):S44

This article is available from: <http://www.biomedcentral.com/1753-6561/1/S1/S44>

© 2007 Xing et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

A family-based association study design is not only able to localize causative genes more precisely than linkage analysis, but it also helps explain the genetic mechanism underlying the trait under study. Therefore, it can be used to follow up an initial linkage scan. For an association study of binary traits in general pedigrees, we propose a logistic mixture model that regresses the trait value on the genotypic values of markers under investigation and other covariates such as environmental factors. We first tested both the validity and power of the new model by simulating nuclear families inheriting a simple Mendelian trait. It is powerful when the correct disease model is specified and shows much loss of power when the dominance of a model is inversely specified, i.e., a dominant model is wrongly specified as recessive or *vice versa*. We then applied the new model to the Genetic Analysis Workshop (GAW) 15 simulation data to test the performance of the model when adjusting for covariates in the case of complex traits. Adjusting for the covariate that interacts with disease loci improves the power to detect association. The simplest version of the model only takes monogenic inheritance into account, but analysis of the GAW simulation data shows that even this simple model can be powerful for complex traits.

Background

Linkage analysis is a useful tool for the initial exploration of complex diseases, but is limited in its ability to localize the loci potentially segregating for disease susceptibility. Association analysis, which directly tests the association between a trait and marker alleles, can more precisely localize causative genes. For this purpose a family-based association study design can be used to follow up an initial linkage scan. Moreover, it can help explain the genetic

mechanism underlying the trait because extended pedigrees provide more genetic information than a random sample consisting of the same number of individuals. Note also that the often-quoted paper by Risch and Merikangas [1] for advocating an association study for detecting genes of modest effect employed a family-based association study design-in particular, using the transmission-disequilibrium test (TDT) [2] – showing it to be

more powerful than a linkage analysis of affected sib pairs.

There have been several ways proposed to conduct an association study of a binary trait using general pedigrees [3,4], and there have also been joint linkage and association models proposed [5,6]. In this paper we propose a logistic mixture model for an association study of binary traits in general pedigrees. We first tested both the validity and power of this new model by simulating nuclear families inheriting a simple Mendelian trait of monogenic inheritance. We describe this initial study briefly before applying the new model to the Genetic Analysis Workshop 15 (GAW15) simulated complex trait data; in particular, we examined the performance of the model when adjusting for covariates in the case of a complex trait, as this could provide information on familial residual correlations due to common environmental sharing.

Methods

Denote by y_i the phenotype ($y \in \{0, 1\}$, where 1 denotes affected and 0 denotes unaffected) and by g_i the genotype of the i^{th} individual in a pedigree of n members. The likelihood for the pedigree data $Y = (y_i)$ is given by a mixture model:

$$L(Y) = \sum_{g_1} \sum_{g_2} \dots \sum_{g_n} \prod_{i=1}^n P(y_i | g_i) P(g_i | g_{i_F}, g_{i_M}),$$

where $P(g_i | g_{i_F}, g_{i_M})$ denotes the conditional probability that individual i has genotype g_i given parental genotypes if he or she is a non-founder, or the probability that individual i has the genotype g_i determined by population genotype frequencies if he or she is a founder; and $P(y_i | g_i)$, the penetrance function, denotes the probability that individual i has phenotype y_i given genotype g_i . The product of these two factors is summed over all possible combinations of genotypes for the pedigree members. Assume a diallelic disease model with alleles D and d , where D is the disease-predisposing allele, i.e., $g_i \in \{DD, Dd, dd\}$. Define the logit of the penetrance function, $\text{logit}[P(y_i = 1 | g_i)]$, to be $\alpha + \beta g_i + \gamma M_i + \lambda X_i$, where M_i denotes the marker to be tested for association and X_i denotes a vector of other covariates such as age and sex. Note that g_i denotes the hypothetical genotype at the putative disease locus, whereas M_i denotes the actual genotype observed for the marker to be tested for association. The constraint $P(y_i = 1 | DD) = P(y_i = 1 | Dd) > P(y_i = 1 | dd)$ corresponds to a dominant model, $P(y_i = 1 | DD) > P(y_i = 1 | Dd) = P(y_i = 1 | dd)$ to a recessive model, and

$$P(y_i = 1 | Dd) = \frac{1}{2} [P(y_i = 1 | DD) + P(y_i = 1 | dd)]$$

to an additive model. The likelihood can be calculated using the Elston-Stewart [7] algorithm in the context of complex segregation analysis. The significance of a marker can be tested by comparing the likelihood with and without this marker in the model logit. Because the finite sample-size null distribution of the likelihood ratio test statistic is not known, to determine the empirical significance level of a particular observed result one can either perform a simulation study based on the null hypothesis by generating unassociated marker data for the sample at hand, or perform a permutation test (e.g., [8]). We classify the current method as model-based because a penetrance function is explicitly specified.

Initial simulation: simple traits

We first describe our initial simulation study, for which we simulated nuclear families consisting of two parents and four children. One diallelic marker was simulated with the minor allele frequency (MAF) $p_D = 0.3$, i.e., a common variant corresponding to the case for which association mapping is advocated [9], and the affection status of all individuals was simulated under nine disease models, covering the relative-risk spectrum from low to high, with the minor allele of this diallelic marker the same as the disease-predisposing allele (Table 1). Among the nine models there were three models simulated under each of dominant, recessive, and additive modes of inheritance. For example, a dominant model with penetrances $(f_0, f_1, f_2) = (0.01, 0.03, 0.03)$ is denoted $D3$: D stands for dominant mode of inheritance, 3 stands for $f_2 = 0.03$, and $\text{logit}(f_j) = \alpha + \beta g_{ij}$, where j denotes the number of copies of the disease allele. Under each model random families were generated and those with at least two affected children were ascertained. In this way we generated 500 replicate samples with 30 families in each sample data set. The diallelic marker and pedigree structures were simulated using the program SimPed [10]. Under the alternative hypothesis of association, the affection status of individuals was simulated according to the penetrance functions of each model; under the null hypothesis of no association, the affection status was randomly simulated according to a disease prevalence given by $(1 - p_D)^2 f_0 + 2p_D(1 - p_D) f_1 + p_D^2 f_2$.

According to an individual's genotype, in particular the number of copies of the disease-predisposing allele, a gen-

Table 1: Parameter settings for simulating disease models

Disease model ^a	f_0^b	f_1^b	f_2^b	p_D^c	Prevalence
D3	0.010	0.030	0.030	0.300	0.020
D6	0.010	0.060	0.060	0.300	0.036
D9	0.010	0.090	0.090	0.300	0.051
R3	0.010	0.010	0.030	0.300	0.012
R6	0.010	0.010	0.060	0.300	0.015
R9	0.010	0.010	0.090	0.300	0.017
A3	0.010	0.020	0.030	0.300	0.016
A6	0.010	0.035	0.060	0.300	0.025
A9	0.010	0.050	0.090	0.300	0.034

^aD, R and A denote (for the minor allele) dominant, recessive, and additive modes of inheritance, respectively.

^b f_j denotes the penetrance of a genotype with j disease-predisposing alleles.

^cDisease-predisposing allele frequency.

otypic value was assigned under the dominant, recessive, or additive model, respectively. For simplicity, we assumed Mendelian transmission probabilities and known disease allele frequencies. We analyzed each data set under three genetic models, i.e., dominant, recessive, and additive models. The significance of association between the trait and the diallelic marker was tested by fitting models with and without the diallelic marker as a covariate and then performing the likelihood-ratio test (LRT). Theoretically, this LRT statistic should asymptotically follow a χ_1^2 distribution, and we report the empirical type I error rate under the null by calculating the percentage of the 500 replicate p -values attaining the nominal level of 0.05. Because the finite sample-size distribution of the LRT statistic is not known, we report a power determined as the percentage of the 500 replicate likelihood ratio statistics under the alternative larger than the cut-off for the top 5% of the 500 replicate LRT statistics under the null. We termed the current method SEGREG, and, as a comparison, we also analyzed the same data sets by another family-based association method [3], which is called ASSOC. The analyses were performed using the programs SEGREG and ASSOC in the Statistical Analysis for Genetic Epidemiology (S.A.G.E.) software suite, Version 5.2 [11].

Results

Simulation results

The two methods had very similar type I error rate and power, though SEGREG showed slightly higher power in most cases (Table 2); therefore we focus below only on the SEGREG results. The empirical type I error rate at a nominal 0.05 significance level was found to be always far below 0.05, regardless of the correctness of the model assumptions. Under the alternative, assuming a correct genetic model always results in more power than assuming a wrong model, which was anticipated. However,

when analyzed under wrong assumptions, the power depends on both the underlying true model and the model assumed for the analysis. If the true disease model is dominant, analysis assuming a recessive model has little power; and if the true disease model is recessive, analysis assuming a dominant model has little power. Analysis assuming an additive model usually has fair power. However, in the case of the low penetrance disease models D3 and R3, wrongly assuming an additive model has power of only 0.530 and 0.450, respectively. Note, however, that even the power on correctly assuming an additive model is only 0.552 when the true model is A3. When the true disease model is additive, analysis assuming a dominant model is usually much more powerful than that assuming a recessive model.

Application to the GAW15 simulated data: complex traits

Compared to our simulation of simple traits, the GAW15 simulation data provided an opportunity to test the new model on a complex trait. In particular, we examined the performance of the model when adjusting for covariates in the case of the data simulating rheumatoid arthritis (RA). Being aware of the answer to the simulated data that smoking interacts with locus B (MAF = 0.35) on chromosome 8 and locus F (MAF = 0.50) on chromosome 11 to increase susceptibility to RA, we performed an association analysis screening for the binary RA trait loci on chromosomes 8 and 11 using all 100 replicates of the simulated 10 K SNP data. The empirical power at a nominal 0.05 significance level for loci B and F was determined by comparing the likelihood ratio statistics to the distribution of statistics for the disease-unassociated markers SNP1_3 (MAF = 0.35) and SNP1_4 (MAF = 0.50), respectively. Age and sex were always included as covariates, and we compared the results with and without adjusting for smoking. To mimic a real situation, we only chose the first 100 nuclear families in each replicate for this study. Markers on chromosome 8 and SNP1_3 were coded with the minor allele dominant and markers on chromosome 11

Table 2: Empirical type I error rate at a nominal 0.05 significance level and power at an estimated 0.05 significance level under various assumed genetic models, based on 500 replicate samples

Simulation model	Analysis model											
	Dominant				Recessive				Additive			
	SEGREG		ASSOC		SEGREG		ASSOC		SEGREG		ASSOC	
	Type I error	Power	Type I error	Power	Type I error	Power	Type I error	Power	Type I error	Power	Type I error	Power
D3	0.006	0.828	0.006	0.818	0.014	0.074	0.010	0.080	0.000	0.530	0.000	0.524
D6	0.016	0.988	0.016	0.986	0.030	0.062	0.018	0.056	0.004	0.852	0.004	0.824
D9	0.004	1.000	0.004	1.000	0.022	0.058	0.022	0.068	0.004	0.940	0.002	0.914
R3	0.004	0.036	0.004	0.036	0.030	0.646	0.024	0.706	0.006	0.450	0.004	0.468
R6	0.008	0.030	0.010	0.028	0.020	0.998	0.022	1.000	0.002	0.978	0.004	0.982
R9	0.016	0.064	0.016	0.056	0.018	1.000	0.018	1.000	0.010	0.996	0.012	1.000
A3	0.012	0.468	0.012	0.462	0.022	0.232	0.020	0.250	0.001	0.552	0.002	0.546
A6	0.008	0.844	0.008	0.838	0.030	0.324	0.022	0.352	0.012	0.874	0.012	0.854
A9	0.010	0.944	0.014	0.936	0.016	0.550	0.010	0.574	0.001	0.962	0.004	0.960

and SNP1_4 were coded in an additive fashion, corresponding to the fact that loci B and F were simulated under dominant and additive models, respectively.

Without adjusting for smoking, the power of detecting loci B and F was 0.33 and 0.45 respectively, whereas the power increased to 0.35 and 0.56 after taking this covariate into consideration.

Discussion

The new method is illustrated here on the full likelihood function of a pedigree, which is usually ascertained according to the phenotypes of probands instead of random sampling, and so, without appropriate ascertainment correction, the parameter estimates may be biased; however, this does not affect validity for testing the significance of association as a generalized linear model, though proper parameter inference can be made only when correcting for ascertainment or building the model on a conditional likelihood [12,13]. The LRT using large sample theory was very conservative for both SEGREG and ASSOC in the current study. We speculate this is because of the small sample size and we hypothesize that, as the sample size increases, the type I error rate will come closer to the nominal level. This topic awaits further investigation.

As a model-based approach, our method requires pre-specifying a genetic model and explicit penetrance functions. According to the simulation study, great power loss occurs when the dominance of a model is inversely specified, i.e., a dominant model is wrongly specified as recessive and *vice versa*, which is similar to the situation in

model-based linkage analysis [14]. This is no surprise, because both are built on the full likelihood function of a pedigree. In general, the results suggest using an additive model in practice. However, in the common variant low penetrance scenario (D3, R3, A3), under the true genetic mechanism both dominant and recessive models have appreciably higher power than an additive model (0.828, 0.646, and 0.552, respectively). The fundamental reason for this lies in the three components of the mixture distributions not being as easily distinguishable as two components in the case of low penetrance. It is of interest to note that ASSOC, as a model-free method in the sense that no penetrance function is specified, shows similar sensitivity to the marker coding scheme as SEGREG. Because most association methods require coding SNPs in a dominant, recessive, or additive fashion, we speculate that this observation is applicable to most methods. Therefore, caution should be taken when coding markers regardless of the method employed to test association.

Detecting variants of modest effect remains a challenge for association studies, especially in the case of rare variants, which were not even simulated in the current study. For the GAW simulation data, the proposed method showed more power in detecting locus F than in detecting locus B. The effect of locus F was simulated via a variance-component method for the continuous trait IgM, which in turn affected the RA trait, whereas the effect of locus B was simulated under a dominant model with relative risk equal to 1.5. There is no direct way to compare their effect sizes. Because only locus F was fairly detected, we speculate that, if measured on the same scale, the effect size of locus B would be much more moderate.

The main limit of our method lies in its assumption of a monogenic disease mechanism without allowing for familial correlation due to polygenic and/or common environmental effects, which is unrealistic for complex diseases (though widely adopted by most methods in the literature). Model-based methods should use models that approximate the complexity of the disease being studied in order to be both robust and powerful. Analyses by models that ignore residual familial correlation can result in decreased power; how to model the familial correlations is a topic for further investigation, though we could easily incorporate covariates that resemble the action of familial correlations into the current model.

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

Some of the results in this paper were obtained by using the program package S.A.G.E., which is supported by a U.S. Public Health Service Resource Grant (RR03655) from the National Center for Research Resources. This work was supported in part by research grants GM28356 from the National Institute of General Medical Sciences and by Cancer Center Support Grant P30 CAD43703 from the National Cancer Institute. CX was supported by a Merck Foundation fellowship. CX and QL were sponsored to attend GAW15 by the Endowment Sponsored Mentorship Program of Case Western Reserve University Graduate School.

This article has been published as part of *BMC Proceedings* Volume 1 Supplement 1, 2007: Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci. The full contents of the supplement are available online at <http://www.biomedcentral.com/1753-6561/1?issue=S1>.

References

1. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273**:1516-1517.
2. Spielman RS, McGinnis RE, Ewens WJ: **Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).** *Am J Hum Genet* 1993, **52**:506-516.
3. Zhu X, Elston RC, Bielefeld RA: **Testing disease-marker association in pedigree data.** In *Proceedings of the Joint Statistical Meeting: 10-14 Aug 1997 Anaheim, CA. Alexandria: American Statistical Association; 1998*:38-43.
4. Rabinowitz D, Laird N: **A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information.** *Hum Hered* 2000, **50**:211-223.
5. Cantor RM, Chen GK, Pajukanta P, Lange K: **Association testing in a linked region using large pedigrees.** *Am J Hum Genet* 2005, **76**:538-542.
6. Li M, Boehnke M, Abecasis GR: **Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal.** *Am J Hum Genet* 2005, **76**:934-949.
7. Elston RC, Stewart J: **A general model for the genetic analysis of pedigree data.** *Hum Hered* 1971, **21**:523-542.
8. Morris AP, Curnow RN, Whittaker JC: **Randomization tests of disease marker associations.** *Ann Hum Genet* 1997, **61**:49-60.
9. Reich DE, Lander ES: **On the allelic spectrum of human disease.** *Trends Genet* 2001, **17**:502-510.
10. Leal SM, Yan K, Muller-Myhsok B: **SimPed: a simulation program to generate haplotype and genotype data for pedigree structures.** *Hum Hered* 2005, **60**:119-122.
11. **S.A.G.E. Statistical analysis for genetic epidemiology. Version 5.2** 2006 [<http://darwin.case.edu/sage>].
12. Risch N: **Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes.** *Am J Hum Genet* 1984, **36**:363-386.
13. Ginsburg E, Malkin I, Elston RC: **Sampling correction in pedigree analysis.** *Stat Appl Genet Mol Biol* 2003, **2**:2.
14. Clerget-Darpoux F, Bonaiti-Pellié C, Hochez J: **Effects of misspecifying genetic parameters in lod score analysis.** *Biometrics* 1986, **42**:393-399.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

