Genomics & Informatics Vol. 5(1) 19-23, March 2007

Regression Models for Haplotype-Based Association Studies

Sohee Oh[1], Junghyun Namkung[2] and Taesung Park[1,2]

[1]Department of Statistics, Seoul National University, Seoul 151-742, Korea, [2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea

Abstract

In this paper, we provide an overview of statistical models for haplotype-based association studies, and summarize their features based on the design matrix. We classify the design matrix into the two types: direct and indirect. For these two kinds of matrices, we present and compare characteristics using a simple hypothetical example, and a real data set. The motivation behind this study was to provide practitioners with an improved understanding, to facilitate the informed selection of the appropriate haplotype-based model and to improve the interpretability of the models.

Key Words: Association, Case-control Study, Design Matrix, Generalized Linear Model, Haplotype

Introduction

Several statistical approaches have been developed for the analysis of association of single nucleotide polymorphisms (SNPs) and disease phenotypes. However, it has been shown that the association tests which rely on SNPs may loose power due to linkage disequilibrium (LD) among the tested SNPs (Epstein and Satten, 2003). To overcome this problem, alternative approaches based on haplotypes have been developed. Haplotypes are specific combinations of alleles at several loci on the same chromosome. They can sometimes provide greater analytical power than single-marker analysis for genetic association studies. This is because haplotypes are inherited together in the majority of cases, and they incorporate linkage disequilibrium information (Akey and Xiong, 2003; Schaid, 2004). Conversely, haplotype-based statistical analysis has a weakness since haplotypes are often not directly observable. Hence, haplotypes and their frequencies are inferred by statistical methods such as the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977; Excoffier and Slatkin, 1995) or the Bayesian method (Stephens *et al.*, 2001; Lin *et al.*, 2002).

Within the framework of the generalized linear model (GLM), the haplotype effect on traits can be statistically described and tested. The model can be expressed as $E(Y) = f^{-1}(X\beta)$, where Y denotes the trait, X represents the haplotypes that are coded into the design matrix, β denotes the effects of haplotype, and f is a function that generalizes the usual linear regression such as logistic regression in the case-control study. Several solutions have been proposed to account for the haplotype ambiguity problem in the general form of a GLM. However, since each method has its own merits, it is not easy for researchers to choose the most appropriate model to use for a given situation. Furthermore, the interpretation of the result is not simple and may vary from method to method. Confusion and misinterpretation of the results can often occur.

In this paper, we review regression models for haplotype based association studies and summarize their features focused on the design matrices. We classify the haplotype design matrices into direct and indirect types, and for these we present and compare characteristics with a simple hypothetical example. We then apply the two types of statistical models to a real data set taken from a genetic study undertaken by the CDC Chronic Fatigue Syndrome Research Group (http://www.cdc.gov/nciod/ diseases/cfs/).

Regression models for haplotypes

In order to apply our statistical analysis, we shall designate Y as a random variable of the binary trait representing the disease status with a realization of either 0 or 1 depending on whether the subject is a control or a genuine case, respectively. Let (h_i, h_j) be a random variable that denotes the pair of haplotypes for each individual, $i=j$ or $i\neq j$. Let $H = \{h_1, h_2, ..., h_p\}$ be a set of haplotypes. Then, the maximum number of possible haplotypes is 2^m, where m is the number of SNPs.

In association studies, the main interest lies in estimating the effects of H on Y. We review the statistical models for the haplotype based association studies and compare their characteristics while focusing on the design

*Corresponding author: E-mail tspark@stats.snu.ac.kr
Tel +82-2-880-8924, Fax +82-2-883-6114
Accepted 13 Feb 2007

matrices and their implications. To explain the differences between the design matrices, we consider a simple hypothetical example of five individuals. Table 1 shows their haplotypes and probabilities. Individuals Y_1, Y_3, and Y_4 have unambiguous haplotypes, while the remaining individuals have ambiguous haplotypes. We group the haplotype design matrices into direct and indirect types. The direct type uses the haplotype probabilities in the design matrix and models *E(Y)* directly in terms of haplotype probabilities, while the indirect type uses the number of copies of haplotypes.

Table 1. A simple example of haplotype pairs and probability

Individual	Haplotype pair	Probability
Y_1	(h_1, h_1)	1
Y_2	(h_1, h_4)	0.2
	(h_2, h_3)	0.8
Y_3	(h_2, h_2)	1
Y_4	(h_2, h_4)	1
Y_5	(h_1, h_2)	0.2
	(h_1, h_4)	0.3
	(h_2, h_3)	0.2
	(h_3, h_4)	0.3

Direct Design Matrix (DDM)

The direct type of design matrix relies on the estimated haplotype probabilities (proportions). The contributions of the haplotypes are weighted using these probabilities, and unambiguous pairs of haplotypes are coded 1 for the homozygous and 0.5 for each of the heterozygous haplotypes. All other haplotypes are coded as 0. The contributions of ambiguous pairs of haplotypes in the design matrix are based on the half of the probabilities of the haplotype pairs that are estimated by the EM algorithm or other statistical inference methods. DDM relates the predictors based on haplotype probabilities with the trait values. Zaykin *et al*. (2002) proposed the Haplotype Trend Regression (HTR) model using the direct design matrix. HTR is a regression model that relates haplotype probabilities with disease phenotypes. In Table 2 we present DDM data using a hypothetical example.

Table 2. Design matrix

Individual	Haplotype pair	Probability	DDM				IDM				
			h_1	h_2	h_3	h_4	h_1	h_2	h_3	h_4	weight
Y_1	(h_1, h_1)	1	1	0	0	0	2	0	0	0	1
Y_2	(h_1, h_4)	0.2	0.1	0.4	0.4	0.1	1	0	0	1	0.2
	(h_2, h_3)	0.8					0	1	1	0	0.8
Y_3	(h_2, h_2)	1	0	1	0	0	0	2	0	0	1
Y_4	(h_2, h_4)	1	0	0.5	0	0.5	0	1	0	1	1
Y_5	(h_1, h_2)	0.2	0.25	0.2	0.25	0.3	1	1	0	0	0.2
	(h_1, h_4)	0.3					1	0	0	1	0.3
	(h_2, h_3)	0.2					0	1	1	0	0.2
	(h_3, h_4)	0.3					0	0	1	1	0.3

Indirect Design Matrix (IDM)

Consider a model which is dependent on the number of copies of haplotypes. The design matrix consists of the haplotype counts and the estimated haplotype pair probabilities. The indirect design matrix haplotype pairs are coded as 2 for the homozygous haplotypes, and 1 for each of the heterozygous haplotypes irrespective of the

Table 3. Comparison of Methods for Haplotype association based on the Design Matrix

	Haplotype presentation	Interpretation			
Model without intercept	Meaning of Significance for haplotype-specific	Haplotypes that increase(decrease) the odds of trait			
	Meaning of Significance for Global haplotypes	All haplotypes have no additive effect on trait			
Model with intercept	Meaning of Significance for haplotype-specific	The odds of trait compare as baseline haplotype			
	Meaning of Significance for Global haplotypes	No difference in effect compare to baseline haplotype			
		Model with intercept			
		DDM	IDM		
Software	Implemented Program	R package - gap HTR R package – haplo.stats haplo.score	R package – haplo.stats Haplo.glm	Chaplin	HPlus
	Statistic	F-statistic Permutation	LR statistic (Global) t-statistic (haplotype-specific)	Robust score and LR statistic(Global) Wald statistic (haplotype-specific)	Score statistic
	Type of Trait	Exponential family	Exponential family	Binomial	Binomial
	Likelihood	Prospective	Prospective	Retrospective	Prospective
	Environmental covariates	No	Yes	No	Yes
	Missing data allowed	Genotypes	Genotype and/or Environmental covariates	Genotypes	Genotypes
	Assumption of HWE	Population	Population	Controls	Controls

haplotype ambiguity. All other haplotypes are coded as 0. The model is fitted by a weighted estimation method using the haplotype pair probability as weight. The model using the indirect design matrix assumes the additive effect of the haplotypes, and is a special case of the weighted GLM. We can also account for dominant or recessive haplotype effects within this framework. This matrix has been used previously in association studies (Lake *et al*., 2003; Epstein and Satten, 2003; Zhao *et al*., 2003). In table 2 we present IDM data using a hypothetical example, and in Table 3 we summarize the proposed features.

Matrices using the 'two type' design can cause problems during statistical modeling. In DDM, the sum of haplotypes for one individual is 1, while in IDM the sum of haplotypes for one row is 2. Since haplotypes are reconstructed from genotype data, there are always two haplotypes. In this case, a multicollinearity problem occurs during model fitting. Multicollinearity refers to any linear relationship among covariates in linear models. In the presence of multicollinearity, the estimation of model parameters becomes unstable. There are two approaches available to resolve this problem. One approach is to fit the model without an intercept. The second approach is to exclude one haplotype in order to reduce the rank of the design matrix. The excluded haplotype is called the baseline haplotype. The effect of other haplotypes represents the relative risk compared to the baseline haplotype. The most frequent haplotype is commonly used as a baseline.

Firstly, let us consider the test based on the models without an intercept. In this case, we can test the individual k^{th} haplotype effect using the null hypothesis $H_0 : \beta_k = 0$. The significance of this individual haplotype implies that those who have this haplotype have a higher risk than those who do not have this haplotype, when $\beta_k > 0$. For the logistic regression model, this risk can be represented by the odds ratios. In addition, we can test the global haplotype effects using the null hypothesis that all elements of β are zero, that is, $H_0 : \beta = 0$ which implies no haplotype effects on the trait.

Next, consider the test based on the model which excludes a baseline haplotype. We can test the individual relative haplotype effect by testing the null hypothesis $H_0 : \beta_k = 0$. The significance of this individual haplotype implies that those who have this haplotype have a higher risk than those who have the baseline haplotype, when $\beta_k > 0$. In addition, we can test for the global effects of haplotypes using the null hypothesis that all elements of β are zero. The significance of the test result implies that there are some differences among the effects of the haplotypes on the trait.

The meaning of the global test for the model without an intercept is different from that for the model with an intercept. Therefore, the global test results need to be interpreted carefully. For the model without an intercept, the test for the hypothesis $H_0 : \beta_1 - \beta_2 = \cdots = \beta_1 - \beta_k$ is equivalent to the global test for the model with an intercept.

The interpretation of the test results for each design matrix is presented in table 3. It also summarizes the list and the important features of the software programs that implement these models. These features include test statistics, trait types, likelihood, the handling ability for environmental covariates and missing observations, and a test for the Hardy-Weinberg equilibrium (HWE).

Example

In this section, we compare characteristics of two design matrices and present the interpretation of haplotype effects using a dataset from the CDC Chronic Fatigue Syndrome Research Group (http://www.cdc.gov/nciod/diseases/cfs/). In this example, we used a CFS group as the case study (55 subjects) and a non-fatigued group as the control (54 subjects). We analyzed three SNPs, rs258750, rs6188, and rs852977 in the NR3C1 gene which are known to associate with CFS (Online Mendelian Inheritance in Man (OMIM)). The estimated haplotype frequencies by the EM algorithm are shown in table 4. We did not consider haplotype 7, which has an estimated frequency of zero.

We fit the logistic regression models, with and without an intercept, using two types of design matrices. Lake *et al*. (2003) mentioned that haplotype frequencies should be at least 5% to avoid biased estimations of the regression parameters. For this reason, the haplotypes with frequencies below 0.05 were pooled together into a rare haplotype group, denoted by h_r. Tables 5 shows the results of the model fit obtained using SAS version 9.1 (SAS Institute Inc., Cary, NC, USA) with both DDM and IDM. For models without an intercept using both matrices, haplotypes 4 and 5 had statistically significant effects at the 5% confidence level, while all other effects were not significant. The global test yielded a significant

Table 4. Estimated haplotypes and frequencies in the NR3C1 gene

No.	Haplotype	Total Frequency	Control Frequency	Case Frequency
1	AAA	0.00979	0.00743	0.01316
2	AAG	0.02393	0.02202	0.02462
3	ACA	0.02134	0.04513	0
4	ACG	0.60094	0.49098	0.70674
5	GAA	0.17528	0.25299	0.09592
6	GAG	0.13962	0.16200	0.12085
7	GCA	0	0	0
8	GCG	0.02909	0.01944	0.03847

Table 5. Results of model fit

1. No intercept model

Parameter	DDM [1]			I DM [1]		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value
Intercept						
h_4 ACG	0.6574	0.3004	0.0286	0.3371	0.1498	0.0245
h_5 GAA	-1.8147	0.7204	0.0118	-0.8797	0.3526	0.0126
h_6 GAG	-0.5810	0.7409	0.4329	-0.3262	0.3641	0.3703
h_r[2]	0.0377	0.9679	0.9689	-0.0299	0.4576	0.9478
Global hypothesis All $\beta = 0$	LR statistic : 11.6366 df : 4		0.0203	LR statistic : 11.6519 df : 4		0.0201
Model Information Criteria (AIC)	147.470			147.454		

Results from [1]SAS HAPLOTYPE and LOGISTIC.

[2]We used a cut-off value for rare haplotypes of 0.05.

2. Intercept model

Parameter	DDM [1]			I DM [1]		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value
Intercept	0.6574	0.3004	0.0286	0.6741	0.2997	0.0245
h_4 ACG						
h_5 GAA	-2.4721	0.8191	0.0025	-1.2167	0.4048	0.0026
h_6 GAG	-1.2384	0.8616	0.1506	-0.6632	0.4241	0.1179
h_r[2]	-0.6197	1.0432	0.5525	-0.3670	0.4931	0.4568
Global hypothesis All $\beta = 0$	LR statistic : 11.6274 df : 3		0.0088	LR statistic : 11.6428 df : 3		0.0087
Model Information Criteria (AIC)	147.470			147.454		

result with 4 degrees of freedom (df), implying that there is at least one significant individual haplotype effect.

For models with an intercept, haplotype 4 was used as a baseline haplotype and in this case only haplotype 5 showed a significant result. The global test yielded a significant result with 3 df, implying that there is at least one individual haplotype that showed a significantly different effect from the baseline haplotype 4. If we are interested in the effects of a specific haplotype on a disease, the models that do not have an intercept should be applied. On the other hand, if we are interested in comparing the relative effects of a haplotype with the baseline haplotype, the models that have an intercept are recommended.

Discussion

In this paper we have discussed issues regarding modeling for haplotype association using different design matrices, and we have presented the features of each method within the GLM framework. We classified the haplotype design matrices into a direct type and an indirect type. Additionally, we considered the models with and without an intercept. The direct type of design matrix uses the haplotype probabilities and models their effect directly, while the indirect type uses the number of copies of haplotypes. We applied these methods to a dataset from the CDC's Chronic Fatigue Syndrome Research Group. The elements of the two design matrices are different, as are the values of the parameter estimate and the standard error. For the logistic regression model, the parameter estimate can be represented by the odds ratios. As a result, the same haplotype has a different odds ratio according to the choice of design matrices. AIC values are the same for both models with and without an intercept. LR statistics also have similar values for both models. However, when the intercept is included in the model, the df of the global test increases. This increase relates to the power of the test; namely, the power of the test is inversely proportional to the numerical value of the degrees of freedom.

Modeling with DDM relies on the estimated proportions of the haplotypes. However, this approach may not account for the variation due to the haplotype estimation. It has been reported that this method may produce biased and inefficient estimation of regression parameters when the effect sizes are large or haplotype uncertainty is high (Lin *et al.*, 2005). DDM does not allow for dominant and recessive effects of haplotypes, while IDM allows for these genetic models. However, it has also been shown that departure from HWE can cause severe biases in modeling with IDM (Satten and Epstein, 2004).

In this paper, we have reviewed and compared the characteristics of commonly used models for haplotype based association studies. We have classified the models into DDM and IDM types. Parameter interpretations of the existence of an intercept were found to be quite different. As shown in Table 3, numerous software applications have been developed to assist with haplotype based association studies. Unfortunately, the output from these programs is not easy to interpret. This can cause misinterpretation among users, and lead them to draw the wrong conclusions about the effects of haplotypes. The main motivation of this study was to provide practioners with a better insight to help them choose an appropriate model, and improve the interpretability of such models. Our empirical studies have demonstrated that the significant haplotypes can differ from model to model depending on the design matrix. Thus, researchers must choose an appropriate model according to their analysis purpose and interpret the results accordingly.

Acknowledgements

This work was supported by the National Research Laboratory Program of the Korea Science and Engineering Foundation (M10500000126).

References

Akey, J. and Xiong, M. (2001). Haplotypes vs Single marker linkage disequilibrium tests: what do we gain? *Eur. J. Hum. Genet*. 9, 291-300.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *J. R. Stat. Soc*. 39, 1-38.

Epstein, M.P. and Satten, G.A. (2003). Inference on Haplotype Effects in Case-Control Studies Using Unphased Genotype Data. *Am. J. Hum. Genet*. 73, 1316-1329.

Excoffier, L. and Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol*. 12, 921-927.

Lake, S.L., Lyon, H., Tantisira, K., Silverman, E.K., Weiss, S.T., Laird, N.M., and Schaid, D.J. (2003). Estimation and Tests of Haplotype-Environment Interaction when Linkage Phase Is Ambiguous. *Hum. Hered*. 55, 56-65.

Lin, D.Y., Zeng, D., and Millikan, R. (2005). Maximum Likelihood Estimation of Haplotype Effects and Haplotype-Environment Interactions in Association Studies. *Genet. Epidemiol*. 29, 299-312.

Lin, S., Cutler, D.J., Zwick, M.E., and Chakravarti, A. (2002). Haplotype Inference in Random Population Samples. *Am. J. Hum. Genet*. 71, 1129-1137.

Satten, G.A. and Epstein, M. (2004). Comparison of Prospective and Retrospective Methods for Haplotype Inference in Case-Control Studies. *Genet. Epidemiol*. 27, 192-201.

Schaid, D.J. (2004). Evaluating Associations of Haplotypes With Traits. *Genet. Epidemiol*. 27, 348-364.

Stephens, M., Smith, N.J., and Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *Am. J. Hum. Genet*. 68, 978-989.

Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J., and Ehm, M.G. (2002). Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals. *Hum. Hered*. 53, 79-91.

Zhao, L.P., Li, S.S., and Khalid, N. (2003). A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am. J. Hum. Genet*. 72, 1231-1250.