

ON A FAMILY-BASED HAPLOTYPE PATTERN MINING METHOD FOR LINKAGE DISEQUILIBRIUM MAPPING

SHUANGLIN ZHANG

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

Department of Mathematical Science, Michigan Technological University, Houghton, MI 49931, USA

Email: shuzhang@mtu.edu

KUI ZHANG

Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA

Email: kuizhang@hto.usc.edu

JINMING LI AND HONGYU ZHAO

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

Email: {jinming.li, hongyu.zhao}@yale.edu

Linkage disequilibrium mapping is an important tool in disease gene mapping. Recently, Toivonen et al. [1] introduced a haplotype mining (HPM) method that is applicable to data consisting of unrelated high-risk and normal haplotypes. The HPM method orders haplotypes by their strength of association with trait values, and uses all haplotypes exceeding a given threshold of strength of association to predict the gene location. In this study, we extend the HPM method to pedigree data by measuring the strength of association between a haplotype and quantitative traits of interest using the Quantitative Pedigree Disequilibrium Test proposed by Zhang et al. [2]. This family-based HPM (F-HPM) method can incorporate haplotype information across a set of markers and allow both missing marker data and ambiguous haplotype information. We use a simulation procedure to evaluate the statistical significance of the patterns identified from the F-HPM method. When the F-HPM method is applied to analyze the sequence data from the seven candidate genes in the simulated data sets in the 12th Genetic Analysis Workshop, the association between genes and traits can be detected with high power, and the estimated locations of the trait loci are close to the true sites.

Key words: Linkage disequilibrium mapping, data mining, quantitative trait, extended pedigree

1 Introduction

Linkage disequilibrium mapping (LDM) is a powerful method for the identification of disease genes. With the completion of the Human Genome Project, many genetic markers can be identified and genotyped within a very short physical distance, and LDM methods that use a set of markers simultaneously through the consideration of haplotypes across a set of markers may be more powerful than the methods that examine each individual marker separately. Various statistical methods have been proposed to locate disease mutation site based on LD around a disease susceptibility (DS) gene [3, 4, 5, 6, 7, 8]. The power of these methods, as well as their ability to identify the correct position of the DS gene, has been shown to be better than the traditional method based on the LD of two markers. However, most of these methods have been developed under explicit assumptions on the mode of inheritance of the disease and the population history of the studied population, and the effects of violations of these assumptions on the analysis of real data are not well understood. Recently, Toivonen et al. [1] proposed haplotype pattern mining (HPM), a technique that uses data mining methods in LD-based gene mapping. The HPM method aims to identify recurrent haplotype patterns and the haplotype patterns are sorted by the strength of their association to the disease. This method, applicable to data consisting of independent high-risk and normal haplotypes, works with a non-parametric statistical model without any genetic model assumption and allows for missing and erroneous markers within the haplotypes. Toivonen et al. [1] showed that the localization power of the method is high, even when the association is weak. However, there are three limitations for the method described by Toivonen and colleagues. First, related individuals cannot be analyzed in the same analysis because their method is only applicable to case-control data. Secondly, the method is only applicable to binary trait. Thirdly, their approach is purely descriptive and the statistical significance of the observed patterns cannot be assessed.

In this article, we introduce a Family-based Haplotype Pattern Mining (F-HPM) method that extends the HPM method. To allow simultaneous use of related individuals with quantitative trait from an extended pedigree, we employ the Quantitative Pedigree Disequilibrium Test (QPDT) statistic [2] to measure the strength of association between a haplotype and a quantitative trait. We then use a simulation method to assess the statistical significance for the observed patterns. When we apply the F-HPM method to analyze the sequence data of the seven candidate genes from the simulated data sets in the 12th Genetic Analysis Workshop (GAW12), the estimated locations of the trait loci are very close to the true sites and the genes having association with certain traits can be detected with high power.

2 Methods

The idea behind the F-HPM method, as well as the HPM method, is that haplotype patterns close to the DS locus are likely to have stronger association than haplotypes further away. Based on pedigree data that includes genotypes at a set of markers and the quantitative traits with possible missing values of the individuals, there are four steps in the F-HPM method: (1) reconstruct each individual's haplotypes across a set of markers and define the haplotype patterns; (2) for each haplotype pattern P , calculate the QPDT statistic [2] to detect if there is a *strong* association between P and a quantitative trait; (3) calculate the proportion of strongly associated haplotypes around a candidate locus L ; and (4) use a simulation procedure to estimate the statistical significance for the observed association. We describe these four steps in detail in the following discussion.

2.1 Haplotype inference and haplotype pattern

Even with large pedigrees, we may not be able to infer the haplotypes of the individuals unambiguously, especially for the case that the haplotypes are across a large number of markers and there are missing genotype data for some individuals in the pedigree. For uncertainties in haplotype inferences, one method would be to estimate the probabilities for all compatible haplotypes. However, such probabilities depend on many parameters related to the population structure under study, as well as the parameters related to the disease model that we usually have little knowledge about. In our haplotype inference, we use the program HAPLORE (unpublished results; <http://bioinformatics.med.yale.edu>) to reconstruct each individual's haplotypes that include possible ambiguous data at certain markers. The algorithms implemented in HAPLORE are similar to those discussed by Wijtsman [9]. Suppose that chromosome region we examine consists k markers, we denote a haplotype of an individual by a vector $H = (b_1, \dots, b_k)$, where b_i is either an allele at marker i if the haplotypes can be reconstructed unambiguously at this marker, or is a symbol "*" if the haplotypes cannot be reconstructed unambiguously.

We examine the association by looking for haplotype pattern that consists of a set of nearby markers, not necessarily consecutive ones. A haplotype pattern P around marker L is defined as a vector $P = (p_{L-l}, \dots, p_{L-1}, p_L, p_{L+1}, \dots, p_{L+r})$, where each p_i is either an allele of the i th marker or the "don't care" (missing symbol) "*", however, the candidate marker L cannot have a missing symbol. A haplotype pattern P occurs in a given haplotype $H = (b_1, \dots, b_k)$ if $1 \leq L-l < L+r \leq k$ and $p_i = b_i$

or $p_i = *$ for all $i, L-l \leq i \leq L+r$. We use three parameters, the candidate marker L , the number of markers included in haplotype pattern N and the maximum number of the markers with missing data or “don’t care” M to control the haplotype pattern. For example, for a given haplotype vector (2, 3, 5, 7, *, 6, 5, 8, 9, 1), all the haplotype patterns with parameters $L=6, N=3$ and $M=2$ that occur in this haplotype are (7, *, 6), (*, *, 6), (*, 6, 5), (*, 6, *), (6, 5, 8), (6, *, 8), (6, 5, *) and (6, *, *).

2.2 Quantitative pedigree disequilibrium test (QPDT)

The QPDT is a TDT type test that allows quantitative traits and arbitrary pedigree structures [2]. The QPDT uses the following three types of nuclear families in an extended pedigree:

- (I) Families with both parents available and at least one parent being heterozygous at the marker being studied.
- (II) Families with one available parent and one or more offspring where all the offspring have the same genotypes.
- (III) Families with at most one available parent and multiple offspring where at least two siblings have different genotypes.

When a haplotype pattern P is studied, we treat P as one allele, denoted by A , and the other haplotype patterns as another allele, denoted by B . Let X_i denote the number of A alleles carried by the i th child and \bar{X} denote the mean number of A alleles among all the offspring in this nuclear family. For the first type of nuclear families, define $X_{im} = 1$ (or -1) if the mother is heterozygous and transmits allele A (or B) to the i th child, and $X_{im} = 0$ if the mother is homozygous. We similarly define X_{if} for the father. For the second type of nuclear families, we only consider offspring-parent pairs with genotypes (BA, BB) or (AA, AB) , and offspring-parent pair with genotypes (BB, BA) or (BA, AA) . The first genotype in the bracket is the offspring's genotype and second genotype in the bracket is the available parent's genotype. We define $X_{(1)} = 1$ if the genotypes for the offspring-parent pair are (AB, BB) or (AA, AB) , $X_{(1)} = -1$ if the genotypes for the parent-offspring pair are (BB, AB) or (AB, AA) , and $X_{(1)} = 0$ for other genotypes of the offspring-parent pair. For the details on the analysis of the second type of nuclear families, see Sun et al. [10, 11]. Define random variables U_1, U_2 , and U_3 as the covariance between the trait values and the genotypes for the first, second, and third types of nuclear families:

$$U_1 = \sum_{i=1}^t (Y_i - \bar{Y})(X_{im} + X_{if}),$$

$$U_2 = \sum_{i=1}^t (Y_i - \bar{Y})X_{(1)}$$

and

$$U_3 = \sum_{i=1}^t (Y_i - \bar{Y})(X_i - \bar{X}),$$

where t is the number of offspring in a nuclear family, and Y_i is the trait value of the i th child for a quantitative trait of interest. Under the null hypothesis of no linkage or no linkage disequilibrium, $E(U_1) = E(U_3) = 0$. However, under null hypothesis of no linkage or no linkage disequilibrium, $E(U_2)$ is equal to 0 under one of the following two conditions:

A1. Males and females with the same genotype at the marker locus have the same mating preference.

A2. Father and mother in each nuclear family are equally likely to be missing given that one parent is missing.

Even if both of the above two assumptions are violated, we can modify U_2 such that $E(U_2) = 0$ under the null hypothesis [11]. In what follows, we assume $E(U_2) = 0$.

For an extended pedigree, let n_1 , n_2 , and n_3 denote the number of the first, second, and third types of nuclear families, respectively. Define

$$D = \frac{1}{n_1 + n_2 + n_3} \left(\sum_{j_1=1}^{n_1} U_{j_1,1} + \sum_{j_2=1}^{n_2} U_{j_2,2} + \sum_{j_3=1}^{n_3} U_{j_3,3} \right),$$

where $U_{j_1,1}$, $U_{j_2,2}$, and $U_{j_3,3}$ are the covariances between the trait values and the genotypes for the j_k th nuclear family of the k th type. For n independent extended pedigrees, let D_l denote the random variable D defined for the l th extended pedigree, then under the null hypothesis of no linkage or no linkage disequilibrium,

$$E(D_l) = 0 \quad (l = 1, 2, \dots, n), \quad \text{Var}\left(\sum_{l=1}^n D_l\right) \approx \sum_{l=1}^n D_l^2 \quad \text{and the test statistic } T = \frac{\sum_{l=1}^n D_l}{\sqrt{\sum_{l=1}^n D_l^2}}$$

is asymptotically normally distributed with mean 0 and variance 1. This test statistic is the QPDT introduced by Zhang et al. [2].

2.3 Measure of degree of association at a locus L

For a marker location L , we measure the degree of association between the haplotypes nearing L and the trait of interest as follows. Let N denote the number of markers we include in a haplotype pattern (including locus L), and M denote the maximum number of missing markers allowed in a haplotype pattern. Define Ω to be the set of haplotype patterns with respect to parameters N , M , and marker location L . We say that haplotype pattern P is “strongly associated” with the trait if $|T| \geq x$, where T is the QPDT statistic and x is an association threshold. In our analysis, we set $x = 1.96$ so that a strong association is approximately equivalent to setting statistical significance level at 5% for each haplotype pattern P . Intuitively, haplotype patterns near the trait locus are likely to have stronger association than haplotype patterns further away from the trait locus. Therefore, the trait locus is likely to be located at the site, where there is a high proportion of strongly associated haplotype patterns [1].

For a given marker L , we compute the frequency of strongly associated haplotype patterns around this marker as

$$f(L) = \frac{\text{The number of strongly associated haplotype patterns in } \Omega}{\text{The number of haplotype patterns in } \Omega}. \quad (1)$$

For each marker L , we use $f(L)$ as a measure of the degree of evidence for association. If we assume that a trait locus exists in the region being examined, we can predict the location of the trait locus to be close to the markers with higher $f(L)$ values. In our analysis, we estimate the trait locus at the marker that gives the largest value of $f(L)$.

2.4 Statistical significance assessment of the observed measure of association

To test the null hypothesis that the region being examined is not associated with the trait of interest, we use $T_{\max} = \max_L f(L)$ as the test statistic for the null hypothesis. We adopt the simulation procedure proposed by Monks and Kaplan [12] to evaluate the statistical significance of the test statistic, and note that simply permuting trait values among the individuals is not a valid procedure. We describe the procedure in the following. For the first type of the nuclear families, under the null hypothesis of

no association, the probability that a heterozygous parent transmits marker allele A and B with equal probabilities. Thus, if the mother is heterozygous, then, X_{im} is equally likely to be 1 and -1 . If there is only one child, then our simulation procedure randomly assign X_{im} as being equal to 1 or -1 with equal probability. Complications arise when more than one child in the family is available. These complications are a result of linkage between the marker and the trait locus. In the presence of linkage, children with shared marker alleles will have similar quantitative traits, even in the absence of association. This can be taken into account by simultaneous randomization of X_{im} (and, similarly, of X_{if}), for heterozygous parents across the sibship. Let $U_{1m} = \sum_{i=1}^t (Y_i - \bar{Y}) X_{im}$ and $U_{1f} = \sum_{i=1}^t (Y_i - \bar{Y}) X_{if}$.

This procedure is equivalent to randomizing the sign of U_{1m} and the sign of U_{1f} with equal probability and then calculate the value $U_1 = U_{1f} + U_{1m}$. Similar procedures are used to simulate genotypes under the null hypothesis for the other two types of families that is equivalent to randomization of the sign of U_2 and the sign of U_3 with equal probability. For each simulated data set, we randomly give the sign of U_{1m} , U_{1f} , U_2 , and U_3 , and recalculate test statistic T and then $f(L)$ and T_{\max} . Note that $U_1 = U_{1m} + U_{1f}$. We can then derive the empirical distribution of T_{\max} based on the calculated test statistics through a set of simulated data sets.

3 Results

3.1 Data Sets

We evaluate the performance of the proposed F-HPM method using the sequence data from the seven candidate genes (G_1, \dots, G_7) in the simulated data sets in GAW12 for the isolated population scenario. Two of the seven candidate genes affect one or two of the five quantitative traits (Q_1, \dots, Q_7). Table 1 summarizes the relationships between the genes and the traits and the sites of the functional alleles. There are multiple functional alleles within G_2 , with changes in either regulatory elements or in the first or second base-pair of a codon, leading to amino acid substitutions. The simulation data set in GAW12 contains 50 replications for the isolated population. For each replication, the data consists of 23 pedigrees with 1497 individuals in total.

Table 1. The relationships between the candidate genes and the quantitative traits.

Gene	Length (kb)	Influence on the quantitative trait(s)	Location(s) of the functional allele(s)
G ₁	20	None	None
G ₂	13	Q ₅	Multiple sites
G ₃	16	None	None
G ₄	20	None	None
G ₅	17	None	None
G ₆	17	Q ₁ and Q ₂	5782
G ₇	20	None	None

3.2 Results on the associations between candidate genes and traits

We apply the F-HPM method to analyze associations between the seven candidate genes and the five quantitative traits. All of the 50 replications of the simulated data sets in the isolated population are used to investigate the false-positive rates and the power of the F-HPM method. Only polymorphic markers whose major allele frequency is less than 95% are used. In the F-HPM method, we set the association threshold at $x = 1.96$, set the maximum number of markers in a haplotype pattern (including marker L) to be 7 ($N = 7$), and allow up to 6 markers with missing information (denoted by “*”) in a haplotype pattern ($M = 6$). For example, the haplotype may contain locus L and 6 markers on either side of L . We vary L from the first polymorphism to the last one in the entire gene, and calculate $f(L)$ for every marker and T_{\max} for the gene. For each gene and each replication, we simulate 200 samples to evaluate the statistical significance of the observed test statistic. The power comparisons between the F-HPM method and two other methods, the QPDT [2] and QST [12] (a score test of linkage for quantitative traits using haplotypes in extended pedigrees and using hierarchical clustering method to group the haplotypes into two groups), based on these 50 replications are summarized in Table 2. Because Q_3 and Q_4 have no associations with any of the genes, they are not shown in Table 2.

According to Table 1, G_1 , G_3 , G_4 , G_5 , and G_7 have no association with all the quantitative traits, G_2 is associated with Q_5 only and G_6 is associated with Q_1 and Q_2 . It can be seen from Table 2 that the false positive rate of our method is within the 95% confidence interval of the nominal level, i.e. 5%. For the power comparison, the QPDT testing one marker at a time has the lowest power in all the cases. For testing the association and linkage between G_1 with single mutation and the traits Q_1 and Q_2 , the two haplotype methods F-HPM and QST have similar

power. However, for detecting G_2 with multiple functional mutations, the F-HPM is more powerful.

Table 2. The power comparisons of the three tests: F-HPM and QPDT (after Bonferroni correction) for the associations and QST for linkage between the seven candidate genes and the five quantitative traits at statistical significance level 5%. There are three true gene-trait associations and the power for these three pairs is denoted in bold face font.

Gene	Q_1			Q_2			Q_5		
	F-HPM	QPDT	QST	F-HPM	QPDT	QST	F-HPM	QPDT	QST
1	0.04	0	0.08	0.10	0	0.14	0.06	0	0.14
2	0.10	0	0.02	0.08	0	0.04	0.68	0.16	0.54
3	0.00	0	0.02	0.02	0	0	0.10	0.02	0.02
4	0.08	0	0	0.02	0	0.10	0.02	0	0
5	0.04	0	0	0.06	0.02	0.02	0.08	0	0.08
6	0.92	0.64	0.98	0.82	0.26	0.80	0.04	0	0.04
7	0.08	0	0.02	0.02	0	0.06	0.08	0	0.02

For every candidate gene, as we vary the marker location along the gene, we obtain a curve of $f(L)$ for each replication. In Figure 1(a), we present the $f(L)$ curves for the association test between G_6 and Q_1 using the first five replicated samples. Although there are variations, the highest peak is near the true site of the functional allele. In Figure 1 (b), we present the $f(L)$ curves for the association test between G_2 and Q_5 . Because there are multiple functional polymorphisms in G_2 , the signal is not as strong as that in Figure 1 (a).

We estimate the trait locus at the marker L with the highest $f(L)$. The histograms for the estimated locations for Q_1 in G_6 and Q_5 in G_2 for those replications in which the trait value has significant associations with the gene are given in Figure 2. For G_6 , the estimated locations of the trait locus for Q_1 are at site 6805 for 32 replications out of 46 significant samples. This estimate is ~ 1 kb from the true location site 5782. For G_6 , the estimated locations of the trait locus for Q_2 have a similar pattern. For G_2 , the estimated locations of the trait locus for Q_5 are in sub-regions around sites 715, 4977, and 12411. The three sites are all within the regulatory regions where the true functional alleles are. Therefore, even there are multiple functional alleles in this gene, the F-HPM method is able to identify the locations of these functional alleles.

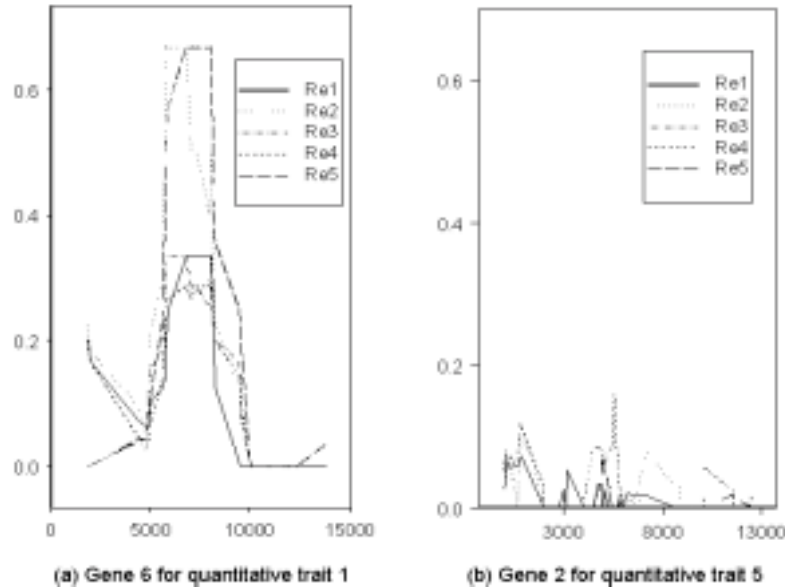


Figure 1. Frequency of strongly associated (with Q_1 for G_6 and with Q_5 for G_2) haplotype patterns versus polymorphic site location.

4 Discussion

We have proposed the F-HPM method to allow simultaneous use of related individuals with quantitative trait from an extended pedigree. This method works with a non-parametric statistical model without any genetic model assumption and allows for missing and erroneous markers within the haplotypes. It tests the association between a set of markers and the quantitative traits and predicts the location of the DS gene at the same time. When we apply the F-HPM method to analyze the sequence data of the seven candidate genes from the simulated data sets in GAW12, the estimated locations of the trait loci are very close to the true sites and the genes having association with certain traits can be detected with higher power comparison with the QPDT [2], the single marker method. For detecting genes with multiple functional mutation, the F-HPM method is more powerful than the QST, another haplotype method.

In the application of the F-HPM method, we need to specify the number of markers included in a haplotype pattern, the number of missing data markers allowed, and the association threshold. The optimal choices of these parameter values need further study, although the method seems to be quite robust with respect to the parameter values for the data analyzed here. From the applications of the F-HPM method to the simulated data sets, we feel that this approach represents a promising method to map complex disease genes.

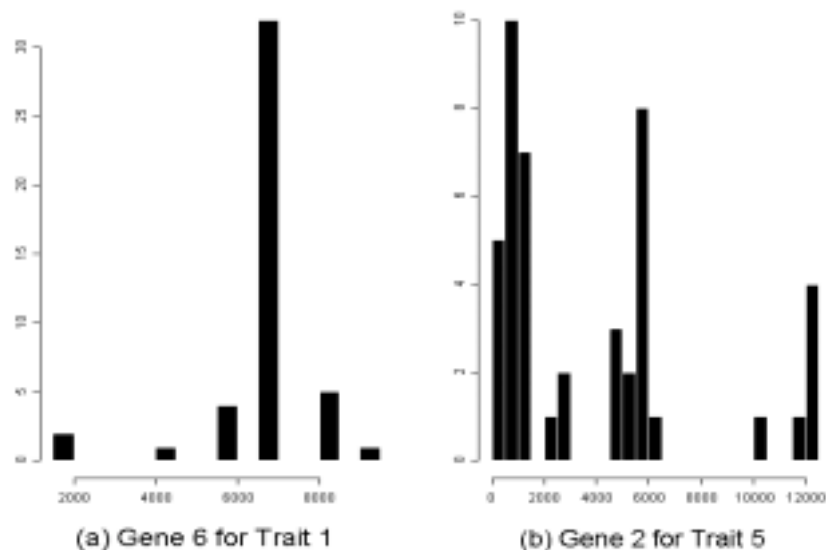


Figure 2. Histograms of estimated locations for Q_1 versus G_6 and Q_5 versus G_2 .

5 Acknowledgements

Supported in part by grants GM59507 and HD36834 from NIH. We thank Dr. MacCluer for providing us the simulated data from GAW12. GAW is supported by NIH grant GM31575 from NIGMS.

References

1. Toivonen H. T. T., Onkamo P., Vasko K., Ollikainen V., Sevon P., Mannila H., Herr M. and Kere J., Data mining applied to linkage disequilibrium. *Am. J. Hum. Genet.* **67** (2000) pp 133–145.

2. Zhang S., Zhang K., Li J., Sun F. and Zhao H., Test of linkage and association for quantitative traits in general pedigree: The quantitative pedigree disequilibrium test. *Genet. Epidemiol.* (2001) in press. Edited by Wijsman E. M., Almasy L., Amos C. I., Borecki I., Falk C. T., King T. M., Martinez M. M., Meyers D., Neuman R., Olson J. M., Rich S., Spence M. A., Thomas D. C., Vieland V. J., Witte J. S. and MacCluer J. W. Analysis of complex genetic traits: Applications to asthma and simulated data.
3. Terwilliger J. D., A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* **56** (1995) pp 777–787.
4. Devlin B., Risch N. and Roeder K., Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36** (1996) pp 1–16.
5. Lazzeroni L. C., Linkage disequilibrium and gene mapping: an empirical least-squares approach. *Am. J. Hum. Genet.* **62** (1998) pp 159–170.
6. McPeck M. S. and Strahs A., Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* **65** (1999) pp 858–875.
7. Service S. K., Temple Lang D. W., Freimer N. B. and Sandkuijl L. A., Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am. J. Hum. Genet.* **64** (1999) pp 1728–1738.
8. Zhang S. and Zhao H., Linkage Disequilibrium Mapping in populations of variable size using the decay of haplotype sharing and a stepwise-mutation model. *Genet. Epidemiol.* **19** (2000) pp S99–S105.
9. Wijsman E. M., A deductive method of haplotype analysis in pedigrees. *Am. J. Hum. Genet.* **41** (1987) pp 356–373.
10. Sun F. Z., Flanders W. D., Yang Q. and Khoury M. J., Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *Am. J. Epidemiol.* **150** (1999) pp 97–104.
11. Sun F. Z., Flanders W. D., Yang Q. and Zhao H., Transmission/disequilibrium test for quantitative traits. *Ann. Hum. Genet.* **64** (2000) pp 555–565.
12. Monks S. A. and Kaplan N. L., Removing the sampling restrictions from family-based test of association for a quantitative-trait locus. *Am. J. Hum. Genet.* **66** (2000) pp 576–592.
13. Li J, Wang D, Dong J, Jiang R, Zhang K, Zhang S, Zhao H, Sun F., The Power of Transmission Disequilibrium Tests for Quantitative. *Genet. Epidemiol.* (2001) in press. Edited by Wijsman E. M., Almasy L., Amos C. I., Borecki I., Falk C. T., King T. M., Martinez M. M., Meyers D., Neuman R., Olson J. M., Rich S., Spence M. A., Thomas D. C., Vieland V. J., Witte J. S. and MacCluer J. W. Analysis of complex genetic traits: Applications to asthma and simulated data.