BMC Genetics

METHODOLOGY ARTICLE Open Access

Statistically efficient association analysis of quantitative traits with haplotypes and untyped SNPs in family studies

Guoqing Diao[1*] and Dan-yu Lin[2]

Abstract

Background: Associations between haplotypes and quantitative traits provide valuable information about the genetic basis of complex human diseases. Haplotypes also provide an effective way to deal with untyped SNPs. Two major challenges arise in haplotype-based association analysis of family data. First, haplotypes may not be inferred with certainty from genotype data. Second, the trait values within a family tend to be correlated because of common genetic and environmental factors.

Results: To address these challenges, we present an efficient likelihood-based approach to analyzing associations of quantitative traits with haplotypes or untyped SNPs. This approach properly accounts for within-family trait correlations and can handle general pedigrees with arbitrary patterns of missing genotypes. We characterize the genetic effects on the quantitative trait by a linear regression model with random effects and develop efficient likelihood-based inference procedures. Extensive simulation studies are conducted to examine the performance of the proposed methods. An application to family data from the Childhood Asthma Management Program Ancillary Genetic Study is provided. A computer program is freely available.

Conclusions: Results from extensive simulation studies show that the proposed methods for testing the haplotype effects on quantitative traits have correct type I error rates and are more powerful than some existing methods.

Keywords: Complex diseases, EM algorithm, Gene-environment interactions, Haplotype analysis, Hardy-Weinberg equilibrium, Unphased genotype, Variance-component models

Background

With the advances in high-throughput genotyping technologies and the availability of dense SNP maps across the human genome [1], haplotype-based association analysis plays an increasingly important role in mapping genes that influence complex human diseases. Haplotypes, which are specific combinations of alleles at several tightly linked SNPs on a chromosome, incorporate the linkage disequilibrium information and pertain to the functional properties of proteins through the amino acids sequences. Association analysis based on haplotypes tends to be more powerful than the analysis of individual SNPs, especially when the causal SNPs are not directly typed or when multiple mutations occur in the *cis* position [2–6].

Standard genotyping procedures only measure unphased genotypes rather than haplotypes. Haplotypes are ambiguous if the genotypes of a subject are heterozygous at more than one marker locus. The ambiguity of the gametic phase information poses a major challenge in the haplotype analysis. For population-based studies with unrelated individuals, a number of methods have been developed to estimate haplotype frequencies or infer indi-

*Correspondence: gdiao@gwu.edu
[1] Department of Biostatistics and Bioinformatics, The George Washington University, Washington, District of Columbia, USA
Full list of author information is available at the end of the article

vidual haplotypes from unphased genotype data [7–12] and to make inference about the effects of haplotypes on disease phenotypes [13–22].

Family studies are more attractive than population-based studies because family data reduce the ambiguity of haplotypes and are less prone to spurious associations caused by population admixture and stratification. Several methods have been developed to estimate haplotype frequencies or infer individual haplotypes from unphased genotype data for general pedigrees, including HAPLORE [23], GENEHUNTER [24], PedPhase [25], and MERLIN [26]. Zhang and Zhao [27] demonstrated through simulation studies that HAPLORE and MERLIN had comparable performance and outperformed the other two methods.

Several methods have been developed for the haplotype association analysis in family studies. Horvath et al. [28] extended the method of Rabinowitz and Laird [29] to multiple markers and proposed a haplotype version of family-based association tests (FBAT). The haplotype FBAT estimates the haplotype frequencies by the expectation-maximization (EM) algorithm [30] under Hardy-Weinberg equilibrium. A score statistic is then constructed in the same manner as the original FBAT except that the genotype score is coded as a weighted sum of haplotype scores, the weight being the conditional probability of a particular haplotype configuration given that it is compatible with the unphased genotype. The haplotype FBAT is computationally simple and can provide either haplotype-specific tests or multi-haplotype tests. However, this method is limited to nuclear families without covariates, does not account for within-family trait correlations, and does not estimate genetic effects. Furthermore, it discards the parental phenotype information and thus may cause substantial loss of power. Dudbridge [31] proposed a retrospective likelihood approach for the association analysis for nuclear families and unrelated subjects with missing genotype data. The retrospective likelihood is based on the probability of observing the parental and offspring genotypes, given the trait values of the all the children in a nuclear family. Other related work includes the family-based association test for dichotomous traits [32], the extension of haplotype FBAT to multiple phenotypes [33], and a Bayesian regression method [34].

Missing genotype data are inevitable in genetic association studies. For example, some study subjects may have missing genotypes at certain SNP loci due to assay failures. Another form of missing data arises when the investigators are interested in untyped SNPs, i.e., the SNPs that are not on the genotyping platform used in the study and thus missing on all study subjects. Haplotypes provide an effective way of inferring the missing genotypes at a particular SNP from the observed genotypes of neighboring SNPs. Lin et al. [35] developed efficient likelihood-based methods to deal with missing genotype data in case-control studies. For family studies, Burdick et al. [36] and Chen and Abecasis [37] imputed the missing genotype values by their expected values via the Elston-Steward or Lander-Green algorithm. Both methods require that at least some members of a family have non-missing genotype data so that they can be used to estimate the conditional distribution of the missing genotypes for other members of the family. These methods cannot be used when the genotype data of the entire family are missing and thus cannot handle untyped SNPs.

In this paper, we present an efficient likelihood-based approach to studying the associations between haplotypes and quantitative traits. This approach estimates the haplotype frequencies and the haplotype effects on the quantitative trait simultaneously. It is very efficient in dealing with missing genotype data. In addition, it allows departures from Hardy-Weinberg equilibrium, accounts for within-family trait correlations, and accommodates general pedigrees with arbitrary patterns of missing data. We characterize the effects of haplotypes on the quantitative trait by a linear regression model with random effects and derive the corresponding likelihood function. We develop efficient likelihood-based estimation and testing procedures. Extensive simulation studies show that the new methods perform well in realistic scenarios. An application to family data from the Childhood Asthma Management Program (CAMP) Ancillary Genetic Study [38] is provided.

Results

Simulation studies

We conducted extensive simulation studies to assess the performance of the new methods in realistic settings. We simulated SNP genotypes according to the haplotype distribution observed in the CEU sample of the HapMap. The inbreeding coefficient ρ was set to 0.02. We generated the quantitative trait values from model (2) in the Methods section with a potentially causal haplotype or SNP. For each scenario, we generated 10,000 data sets, each of which contains 100 nuclear families with two parents and two children.

In the first set of simulation studies, we evaluated the performance of the new method for haplotype association analysis. We were particularly interested in SNPs 20-24 on chromosome 18 of the CEU sample in the HapMap genomewide data. This set of SNPs was previously considered by Lin et al. [35]. The LD among the 5 SNPs is not particularly strong. The five most common haplotypes are 00000, 00011, 00100, 01000, and 01011, with frequencies 0.1431, 0.1312, 0.1941, 0.1756, and 0.1579, respectively.

We generated trait values from the additive model

$$Y_{ij} = \alpha + \beta_1\{I(H_{ij1} = h^*) + I(H_{ij2} = h^*)\} + \beta_2 X_{ij}$$

$$+\beta_3\{I(H_{ij1} = h^*) + I(H_{ij2} = h^*)\}X_{ij} + g_{ij} + e_{ij}, \qquad (1)$$

where the target haplotype h^* is 00100 and the environmental variable X_{ij} is a Bernoulli random variable with 0.3 success probability. The parameters β_1, β_2, and β_3 correspond to the effect of the target haplotype, the effect of the environmental variable, and the haplotype-environment interaction, respectively. We set α, σ_g^2, and σ_e^2 to 1, 0.5, and 1.0, respectively. For making inference on β_1, we set $\beta_3 = 0.4$ and varied β_1 from 0 to 0.4; for making inference on β_3, we set $\beta_1 = 0.4$ and varied β_3 from 0 to 0.4. For each setting, we considered both the situation of no missing genotypes and the situation with 10% randomly missing genotypes.

Tables 1 and 2 summarize the results for estimating the haplotype effect and haplotype-environment interaction, respectively, while Table 3 presents the results for estimating the haplotype frequencies under $\beta_1 = \beta_3 = 0.4$. The estimators of the haplotype effect and haplotype-environment interaction are virtually unbiased, and so are the estimators of the haplotype frequencies. The variance estimators accurately reflect the true variations of the parameter estimators, and the confidence intervals have correct coverage probabilities. The results of the haplotype frequencies estimates are similar to those obtained from HAPLORE. However, our main objective is to conduct haplotype association analysis and the proposed full information maximum likelihood approach typically yields statistically efficient parameter estimators by the parametric likelihood theory.

We also compared the new method to the Haplotype FBAT. Since the latter cannot handle covariates, we set $\beta_2 = \beta_3 = 0$ in model (1). Figures 1 and 2 display the type I error and power of the association tests for the haplotype effect at the nominal significance level of 0.01 without missing data and with 10% missing data, respectively. The new method has the correct type I error and is more powerful than the Haplotype FBAT. The power differences are particularly strong when parental phenotype data are available. The power gain of the new method over the Haplotype FBAT is expected to be even more substantial in the presence of covariate effects. Without missing data, the new method has almost the same power as the ideal case of known haplotypes. The loss of power for the new method caused by missing genotypes is rather moderate, even when there is substantial missingness. These results suggest that the new method can effectively infer the haplotype configuration and is efficient in dealing with missing genotype data.

We next studied the problem of missing genotype data. We considered the same model as before but set SNP 20 to be causal with an additive effect. We let the genotypes of the 5 SNPs be missing independently with a 10% missing

Table 1 Summary statistics for the estimation of the haplotype effect

Effect size	Missing rate	Bias	SE	SEE	CP
0	0.0	0.001	0.137	0.136	0.949
	0.1	0.000	0.138	0.139	0.949
0.1	0.0	0.001	0.136	0.136	0.950
	0.1	-0.003	0.140	0.139	0.949
0.2	0.0	-0.002	0.136	0.136	0.950
	0.1	-0.002	0.139	0.139	0.949
0.3	0.0	0.002	0.133	0.136	0.955
	0.1	-0.002	0.139	0.139	0.946
0.4	0.0	-0.002	0.136	0.136	0.950
	0.1	-0.002	0.139	0.139	0.945

SE is the sampling standard error of the parameter estimator, SEE is the mean of the standard error estimator, and CP is the coverage probability of the 95% confidence interval

Table 2 Summary statistics for the estimation of the haplotype-environment interaction

Effect	Missing				
size	rate	Bias	SE	SEE	CP
0	0.0	0.003	0.237	0.234	0.946
	0.1	0.002	0.241	0.238	0.942
0.1	0.0	0.003	0.237	0.234	0.946
	0.1	0.003	0.241	0.238	0.942
0.2	0.0	0.003	0.237	0.234	0.947
	0.1	0.006	0.241	0.238	0.946
0.3	0.0	0.003	0.237	0.234	0.947
	0.1	0.003	0.241	0.238	0.946
0.4	0.0	0.003	0.237	0.234	0.947
	0.1	0.003	0.241	0.238	0.946

SE is the sampling standard error of the parameter estimator, SEE is the mean of the standard error estimator, and CP is the coverage probability of the 95% confidence interval

Table 3 Summary statistics for the estimation of haplotype frequencies

Haplotype	Bias	SE	SEE	CP
00000	-0.0004	0.0181	0.0178	0.943
00011	0.0003	0.0173	0.0172	0.945
00100	0.0000	0.0202	0.0200	0.947
01000	0.0001	0.0197	0.0193	0.943
01011	-0.0003	0.0186	0.0185	0.945

SE is the sampling standard error of the parameter estimator, SEE is the mean of the standard error estimator, and CP is the coverage probability of the 95% confidence interval

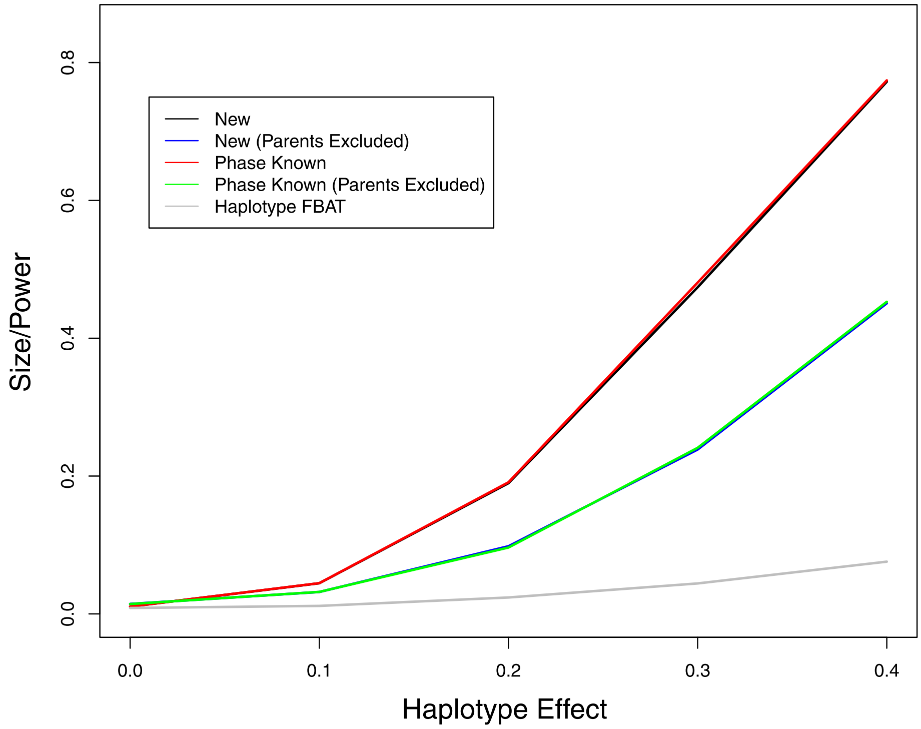

Fig. 1 Type I error and power of association tests on haplotype 00100 at the 1% nominal significance level when there are no missing genotype data

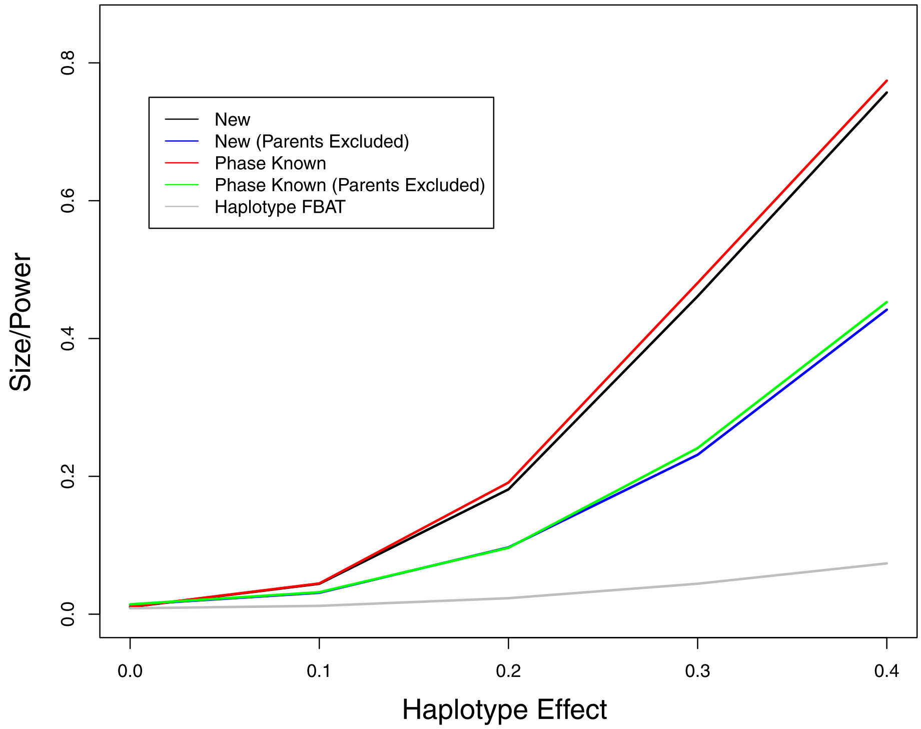

Fig. 2 Type I error and power of association tests on haplotype 00100 at the 1% nominal significance level when there are 10% missing genotype data

rate and performed multi-SNP analysis by including all 5 SNPs in the model. Figure 3 displays the type I error of the association tests at SNP 21, which is null, and the power of the association tests at SNP 20, which is causal. The new method provides accurate control of the type I error. The improvement of the new method over the complete-case analysis is substantial. Compared to the full-data analysis, the new method has little loss of power. We also performed single-SNP analysis by including only the causal SNP in the model and compared the new method to the imputation method by Chen and Abecasis [37]. Figures 4 and 5 show the size/power curves of the association tests at SNP 20 with missing genotype rates of 10% and 20%, respectively. The new method is substantially more powerful than the imputation approach, especially when the missing rate is high.

We finally studied the problem of untyped SNPs. We considered the same model as in the above simulation studies with missing genotype data. We set the causal SNP 20 to be untyped and performed single-SNP analysis on SNP 20. In addition to the study sample, we generated a reference panel with 30 or 60 nuclear families. As shown in Fig. 6, the new method has proper type I error and reasonable power compared to the ideal full-data analysis. The reference panel of 30 families is almost as informative as that of 60 families.

CAMP study

We used the new method to study associations between asthma phenotypes and eight SNPs in the Beta2-Adrenergic Receptor (β_2AR) with data from the CAMP Ancillary Genetic Study and compared the results to those of the haplotype FBAT. The CAMP study was a clinical trial of asthmatic children (mild to moderate asthma) who were randomized to three different treatments. The CAMP data set consists of 2,011 individuals in 652 nuclear families. Four percent of the genotypes at eight β_2AR SNPs were missing. Polymorphisms in β_2AR were found to be associated with several asthma phenotypes in previous studies [39, 40]. In this paper, we considered the standardized mean asthma symptom score. Only 573 individuals had non-missing asthma symptom score data, but genotype data for all individuals were used in the model to infer the haplotype configurations. The same data set was previously analyzed in [28].

As shown in Table 4, the new method and the haplotype FBAT identified the same eight haplotypes with frequencies greater than 0.01. The inbreeding coefficient

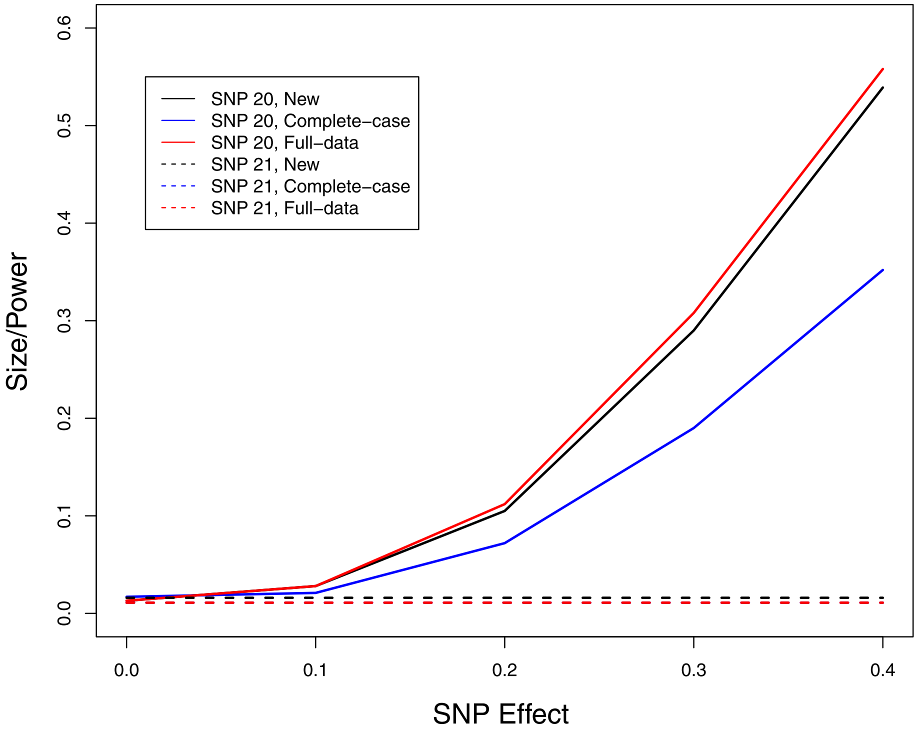

Fig. 3 Type I error and power of association tests at SNP 20, which has a causal additive effect on the phenotype, and SNP 21, which is null, at the 1% nominal significance level when there are 10% missing genotype data. For complete-data analysis, all subjects with missing data are removed. For full-data analysis, the missing genotypes are replaced by their true values

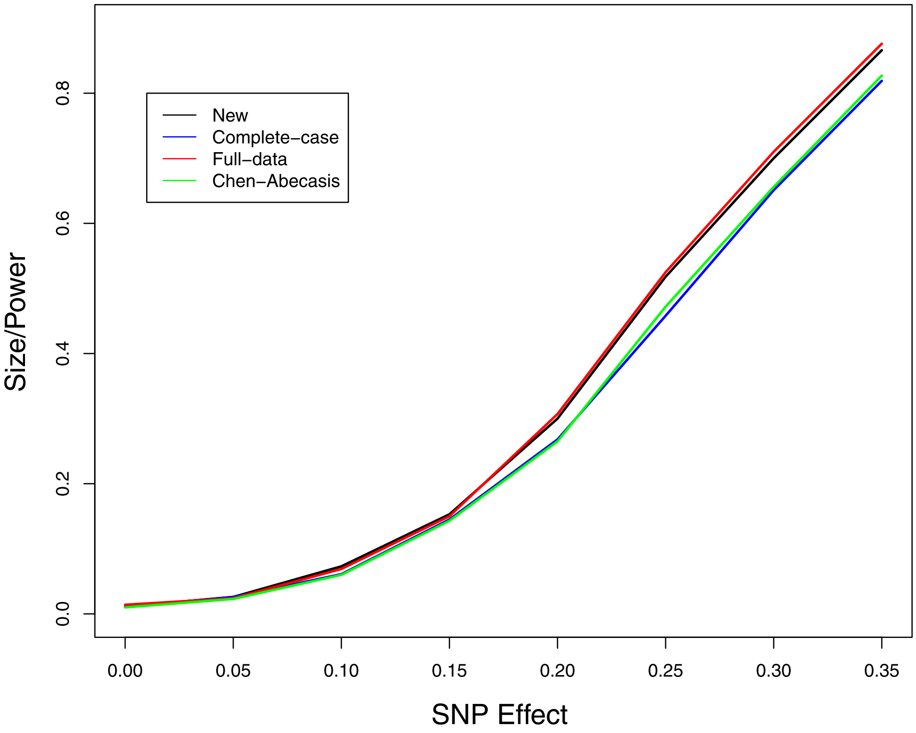

Fig. 4 Type I error and power of single-SNP association tests at SNP 20, which has a causal additive effect on the phenotype, at the 1% nominal significance level when there are 10% missing genotype data. For complete-data analysis, all subjects with missing data are removed. For full-data analysis, the missing genotypes are replaced by their true values

was estimated at 0.04, with a p-value 0.005. The Haplotype FBAT did not detect any significant associations, whereas the new method detected significant associations of haplotypes “12211211” and “11211211” with the mean asthma symptom score, the p-values being 0.017 and 0.023, respectively. Note that the Haplotype FBAT does not estimate haplotype effects.

The results presented in Table 4 pertain to the comparison of a target haplotype with all other haplotypes. We also performed an overall test by comparing six most frequent haplotypes to all other haplotypes within the same model. The resultant LR was 11.66 with 6 degrees of freedom, the corresponding p-value being 0.07. The chi-square test statistic from the haplotype FBAT was 6.98, with a p-value of 0.32.

Conclusion

Haplotype-based association analysis of quantitative traits in family studies is an important tool to identify genes that influence complex human diseases. The existing methods have severe limitations. In this article, we provide an efficient likelihood-based approach to investigating the associations between haplotypes and quantitative traits. Our approach acknowledges the ambiguities of haplotypes in the association analysis by integrating the construction of haplotypes and the estimation of haplotype effects into a single likelihood framework. In addition, our approach accommodates environmental factors, properly accounts for familiar correlations of trait values, and allows departures from Hardy-Weinberg equilibrium.

The proposed method appears to be more powerful than Haplotype FBAT, which is a conditional test. The haplotype FBAT, however, would be more robust to population stratification than the proposed method. To control for spurious association due to population stratification, one may partition the haplotype effect into between- and within-family components, as in [41, 42]. The between-family component accounts for all the spurious association and the within-family component provides a direct measure of the haplotype effect.

In this paper, we use the algorithm described in Zhang et al. [23] to identify the set of all possible haplotype configurations compatible with the observed genotype data. This set can be large for large pedigrees or when the number of SNPs under consideration is large. Although our theory applies to arbitrarily large pedigrees and large number of SNPs, the proposed method is computationally fast and numerically stable when both the pedigrees and the number of SNPs are small. To overcome the computational limitation, one can use the alternative haplotype

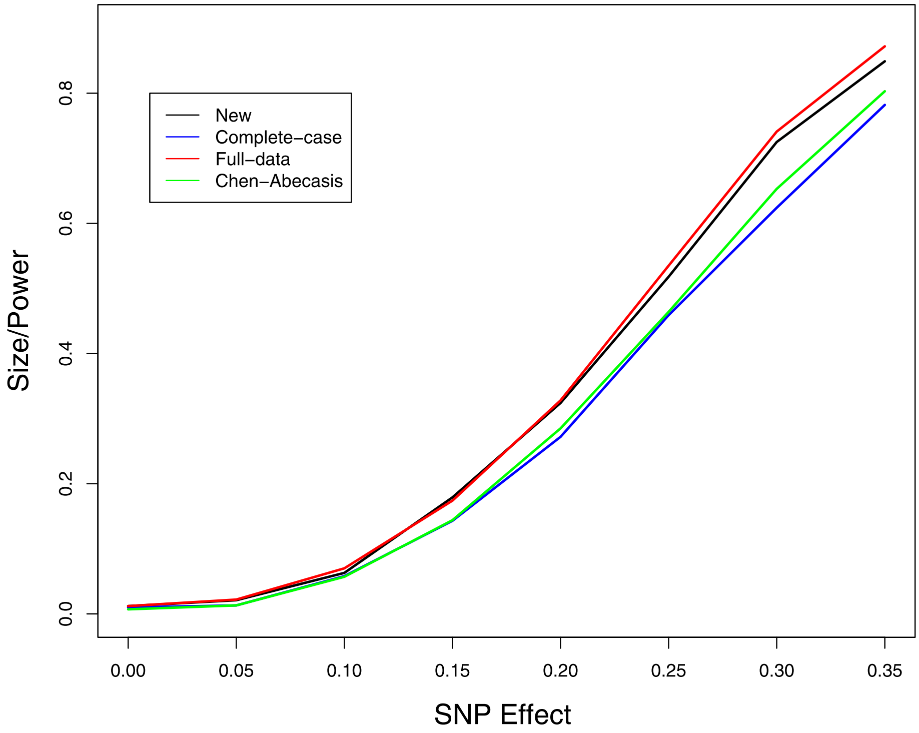

Fig. 5 Type I error and power of single-SNP association tests at SNP 20, which has a causal additive effect on the phenotype, at the 1% nominal significance level when there are 20% missing genotype data. For complete-data analysis, all subjects with missing data are removed. For full-data analysis, the missing genotypes are replaced by their true values

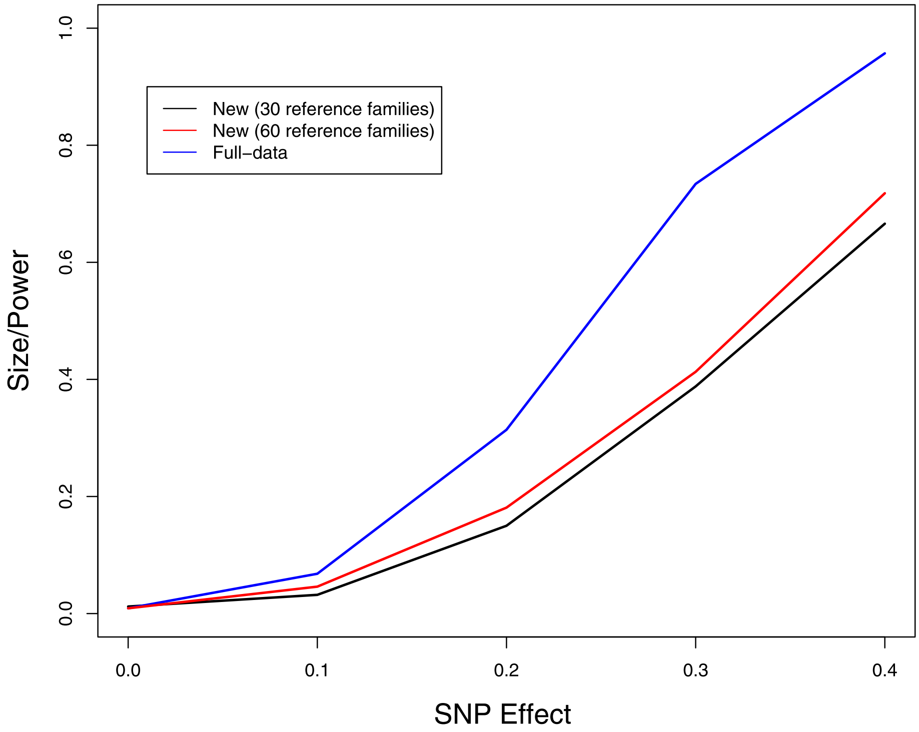

Fig. 6 Type I error and power of association tests of untyped SNP 20 at the 1% nominal significance level. For full-data analysis, the missing genotypes are replaced by their true values

Table 4 Haplotype-specific association analysis of the asthma symptom score in the CAMP study

New Method					Haplotype FBAT		
Haplotype	Freq.	Effect	LRT	*p*-value	Freq.	Z-stat	*p*-value
11122211	0.369	-0.052	0.739	0.390	0.357	0.904	0.366
12211211	0.345	0.144	5.743	0.017	0.352	-0.308	0.758
11221112	0.183	-0.034	0.182	0.669	0.180	-0.879	0.379
11211211	0.038	-0.377	5.153	0.023	0.037	-1.525	0.127
11221111	0.032	-0.052	0.056	0.813	0.029	1.432	0.152
11221122	0.011	0.223	0.782	0.376	0.011	0.998	0.318

reconstruction programs such as MERLIN, which only computes the most likely haplotype configuration. Future research is warranted to compare the statistical efficiency and computational efficiency of different haplotype reconstruction programs.

Some studies involve both families and unrelated individuals. Epstein et al. [43] proposed a likelihood-based approach to single-marker association analysis of binary traits using data from triads and unrelated subjects. For the haplotype-based association analysis of quantitative traits, our approach is applicable to arbitrary pedigrees and therefore can combine information from families and unrelated individuals in a single combined analysis. In fact, we can treat unrelated individuals as unrelated families but with just one individual in each family. This nice feature allows us to extract all available information and further improve the power of association tests.

Under model (2), the quantitative traits within a family follow a multivariate normal distribution. This model may not be appropriate for non-normal traits or traits with outliers. We are currently developing robust semiparametric variance-components models [44] for studying associations between haplotypes and non-normally distributed quantitative traits.

It is desirable to adjust for the effects of multiple testing when considering several haplotype configurations in the same study, especially in genomewide studies. The Bonferroni correction would be overly conservative and permutation would be computationally intensive. Huang et al. [45] proposed an efficient Monte-Carlo approach to adjusting for multiple testing for the haplotype analysis in case-control studies. It would be worthwhile to extend their approach to family studies.

In a haplotype association analysis, the set of all possible haplotype configurations can be large, even for a moderately large number of SNPs. Consequently, the number of parameters included in the model can be huge if we include all possible haplotype configurations in the phenotype model, and numerical computations may not be stable. We suggest specifying one or a few haplotypes of interest in the association test. We may consider a two-step procedure to identify a set of “risk" haplotypes. In the first step, for each possible haplotype configuration with the frequency above a certain threshold (e.g., 0.01), we fit the proposed model and estimate the haplotype effect. In the second step, we include those haplotypes with significant effects (e.g., with p-values <0.05) in the phenotype model. This procedure is similar to some methods in the variable selection literature. It would be interesting to investigate the properties of such a procedure in the future.

In the presence of missing genotype data, one can carry out the single-SNP analysis by using the imputation method, in particular, the multiple imputation procedure [46]. In contrast, the proposed method is based on the full information maximum likelihood, which tends to have more efficient parameter estimators than the multiple imputation method when the model assumptions are satisfied. Additionally, multiple imputation method requires repeated runs of the model, and special care is needed to estimate the standard errors of the parameter estimators. On the other hand, the multiple imputation method is more flexible than the full information maximum likelihood approach.

One limitation of the proposed method is that it is developed for common variant analysis. With the availability of high-throughput sequencing data, it would be interesting to develop rare variance analysis methods for family studies with missing data. This is a topic for future research.

Methods

Suppose that the study contains n families or general pedigrees, with n_i individuals in the ith pedigree. Let Y_{ij} be the quantitative trait of interest and $\mathbf{x}_{ij}$ be a set of environmental variables for the jth member of the ith pedigree. Assume that each individual is genotyped at M tightly linked diallelic SNPs. At each SNP locus, the two possible alleles are denoted by 0 and 1. The total number of possible haplotypes is $K = 2^M$. For example, the possible haplotypes for three SNPs are 000, 001, 010, 011, 100, 101, 110, and 111. For $k = 1, \cdots, K$, let h_k denote the kth possible haplotype. The distribution of the diplotype (i.e., the pair of haplotypes on the two homologous chromosomes) is often assumed to satisfy Hardy-Weinberg equilibrium such that

$$\pi_{kl} = \pi_k \pi_l, k, l = 1, \cdots, K,$$

where π_{kl} is the probability that the diplotype H consists of h_k and h_l, and π_k is the population frequency of haplotype h_k. We consider the following extension of Hardy-Weinberg equilibrium:

$$\pi_{kl} = \begin{cases} \pi_k^2/(1-\rho+\rho\sum_{j=1}^K \pi_j^2), & k = l, \\ (1-\rho)\pi_k\pi_l/(1-\rho+\rho\sum_{j=1}^K \pi_j^2), & k \neq l, \end{cases}$$

where $0 \leq \pi_k \leq 1$, $\sum_{k=1}^K \pi_k = 1$, and ρ is the inbreeding coefficient [47, p. 93]. Excessive homozygosity and excessive heterozygosity arise under $\rho > 0$ and $\rho < 0$, respectively. The special case of $\rho = 0$ corresponds to Hardy-Weinberg equilibrium.

For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, let $H_{ij} \equiv (H_{ij1}, H_{ij2})$ denote the diplotype of the jth member of the ith family, and G_{ij} denote the corresponding multi-locus genotype. Note that G_{ij} codes the number of the "1" allele at each locus such that $G_{ij} = H_{ij1} + H_{ij2}$. We cannot determine H_{ij} from G_{ij} with certainty if the individual is heterozygous at more than one SNP site or if any SNP genotype is missing.

In association studies, we are interested in estimating the effects of H_{ij} and $\mathbf{x}_{ij}$ and possibly their interactions on Y_{ij}. However, we observe G_{ij} instead of H_{ij}. We denote the probability distribution of H_{ij} by $P(H_{ij}; \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ consists of π_k $(k = 1, \cdots, K)$ and ρ. By using a haplotype reconstruction program, such as HAPLORE [23], we can identify the set of all possible haplotype configurations, denoted by $S(\mathbf{G}_i)$, which are compatible with (possibly missing) genotype $\mathbf{G}_i \equiv (G_{i1}, \cdots, G_{in_i})$.

We specify the following linear regression model with random effects

$$Y_{ij} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}(H_{ij1}, H_{ij2}, \mathbf{x}_{ij}) + g_{ij} + e_{ij}, \tag{2}$$

where α is the intercept, $\mathbf{Z}(H_{ij1}, H_{ij2}, \mathbf{x}_{ij})$ is a vector function of (H_{ij1}, H_{ij2}) and $\mathbf{x}_{ij}$, $\boldsymbol{\beta}$ is the corresponding set of regression parameters, g_{ij} is a random effect due to genes at unlinked loci, and e_{ij} is an individual-specific residual environmental effect. Note that g_{ij} is used to capture the correlations of the quantitative trait values within the family. The random variables g_{ij} and e_{ij} are assumed to be independent zero-mean normal with variances σ_g^2 and σ_e^2, respectively. The phenotypic covariance matrix of $\mathbf{Y}_i \equiv (Y_{i1}, \cdots, Y_{in_i})^{\mathrm{T}}$ can be expressed as

$$\mathbf{V}_i = 2\sigma_g^2\boldsymbol{\Sigma}_{gi} + \sigma_e^2\mathbf{I}_i,$$

where $\boldsymbol{\Sigma}_{gi}$ is the matrix of kinship coefficients, and $\mathbf{I}_i$ is an identity matrix.

We define $\mathbf{Z}(H_{ij1}, H_{ij2}, \mathbf{x}_{ij})$ according to the genetic mode of inheritance. For example, the choice of

$$\mathbf{Z}(H_{ij1}, H_{ij2}, \mathbf{x}_{ij}) = \begin{bmatrix} I(H_{ij1} = h^*) + I(H_{ij2} = h^*) \\ \mathbf{x}_{ij} \\ \{I(H_{ij1} = h^*) + I(H_{ij2} = h^*)\}\mathbf{x}_{ij} \end{bmatrix}$$

corresponds to an additive model for the haplotype effect, environmental effects and haplotype-environment interactions, where h^* is the target haplotype of interest, and $I(\cdot)$ is the indicator function. If we are interested in the recessive or dominant effect of h^*, then we set the genotype score in $\mathbf{Z}(H_{ij1}, H_{ij2}, \mathbf{x}_{ij})$ to $I(H_{ij1} = H_{ij2} = h^*)$ or $I(H_{ij1} = h^*$ or $H_{ij2} = h^*)$, respectively. We may include additional terms in $\mathbf{Z}$ so as to assess the effects of several haplotype configurations and to test for multi-haplotype effects.

Write $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \sigma_g^2, \sigma_e^2)^{\mathrm{T}}$. Let $P(\mathbf{Y}_i|\mathbf{x}_i, \mathbf{H}_i; \boldsymbol{\theta})$ denote the multivariate normal density function of $\mathbf{Y}_i$ conditional on $\mathbf{x}_i$ and $\mathbf{H}_i$, where $\mathbf{x}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i})$, and $\mathbf{H}_i = (H_{i1}, \cdots, H_{in_i})$. Let $P(\mathbf{H}_i; \boldsymbol{\gamma})$ denote the probability distribution of $\mathbf{H}_i$. Note that $P(\mathbf{H}_i; \boldsymbol{\gamma})$ is proportional to

$$\prod_{j \in \{\text{founders}\}} P(H_{ij}; \boldsymbol{\gamma}).$$

The complete-data likelihood function for parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ given $(\mathbf{Y}_i, \mathbf{x}_i, \mathbf{H}_i, \mathbf{G}_i)$ $(i = 1, \cdots, n)$ is proportional to

$$\prod_{i=1}^n P(\mathbf{Y}_i|\mathbf{x}_i, \mathbf{H}_i; \boldsymbol{\theta})P(\mathbf{H}_i; \boldsymbol{\gamma}),$$

and the likelihood function based on the observed data $(\mathbf{Y}_i, \mathbf{x}_i, \mathbf{G}_i)$ $(i = 1, \cdots, n)$ is proportional to

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^n \sum_{\mathbf{H}_i \in S(\mathbf{G}_i)} P(\mathbf{Y}_i|\mathbf{x}_i, \mathbf{H}_i; \boldsymbol{\theta})P(\mathbf{H}_i; \boldsymbol{\gamma}). \tag{3}$$

It is worth noting, although we use the algorithm in HAPLORE to identify the set $S(\mathbf{G}_i)$, we maximize the likelihood in (3) to obtain the estimators of all unknown parameters simultaneously, including haplotype frequencies, regression coefficients, and variance parameters.

Treating the H_{ij}'s as missing data, we maximize (3) through the EM algorithm. In the $(t+1)$th E-step, we calculate the conditional expectation of the logarithm of (3) given the observed data and current parameter estimates $\widehat{\boldsymbol{\theta}}^{(t)}$ and $\widehat{\boldsymbol{\gamma}}^{(t)}$ as follows:

$$\sum_{i=1}^n \sum_{\mathbf{H}_i \in S(\mathbf{G}_i)} \omega_{\mathbf{H}_i}^{(t)} \left\{\log P(\mathbf{Y}_i|\mathbf{x}_i, \mathbf{H}_i; \boldsymbol{\theta}) + \log P(\mathbf{H}_i; \boldsymbol{\gamma})\right\}, \tag{4}$$

where

$$\omega_{\mathbf{H}_i}^{(t)} = \frac{P(\mathbf{Y}_i|\mathbf{x}_i, \mathbf{H}_i; \widehat{\boldsymbol{\theta}}^{(t)})P(\mathbf{H}_i; \widehat{\boldsymbol{\gamma}}^{(t)})}{\sum_{\mathbf{H}_i^* \in S(\mathbf{G}_i)} P(\mathbf{Y}_i|\mathbf{x}_i, \mathbf{H}_i^*; \widehat{\boldsymbol{\theta}}^{(t)})P(\mathbf{H}_i^*; \widehat{\boldsymbol{\gamma}}^{(t)})},$$

which is the conditional probability that the haplotype configuration for the ith family is $\mathbf{H}_i$ given the observed data and current parameter estimates. In the $(t+1)$th M-step, we maximize (4) to update parameter estimates. We iterate the E-step and the M-step until convergence. One can also maximize the observed-data likelihood function (3) directly by using an optimization algorithm [48]. The resultant maximum likelihood estimator (MLE) is denoted by $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$. By extending the arguments of Lin and Zeng [21] and Diao and Lin [44], we can show that

$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$ is consistent, asymptotically normal and asymptotically efficient. In addition, the covariance matrix of $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$ can be estimated by the inverse of the observed information matrix. To test the haplotype effects or haplotype-environment interactions, we calculate the likelihood ratio test statistic

$$LR = -2[\log L(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}}) - \log L(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})],$$

where $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}})$ is the restricted MLE under the null hypothesis. The null distribution of LR is asymptotically χ^2 with the degrees of freedom equal to the number of the regression parameters in the null hypothesis. (Note that we are testing the regression effects, not the variance components.) The Wald statistic can also be used to perform hypothesis testing and construct confidence intervals.

Analysis of single SNPs with missing genotypes can be treated as a special case of the proposed haplotype analysis. If we are interested in the additive effect of a particular SNP, then we set the genotype score in $Z(H_{ij1}, H_{ij2}, \mathbf{x}_{ij})$ in model (2) to be the value of $(H_{ij1} + H_{ij2})$ at that SNP position; recessive and dominant effects can be similarly modeled. We can also define $Z(H_{ij1}, H_{ij2}, \mathbf{x}_{ij})$ to formulate the joint effects of all M SNPs or any subset of them.

When one of the M SNPs is untyped, there is no information in the study data to estimate the joint distribution of the M SNPs. We can infer the joint distribution from an external reference database, such as the HapMap. Naturally, the family study and the reference panel are assumed to come from a common population. Let $L_R(\boldsymbol{\gamma})$ denote the likelihood for $\boldsymbol{\gamma}$ based on the reference database. Then the likelihood for $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ that combines the study data and reference database is

$$L_C(\boldsymbol{\theta}, \boldsymbol{\gamma}) = L(\boldsymbol{\theta}, \boldsymbol{\gamma}) L_R(\boldsymbol{\gamma}).$$

The EM algorithm described earlier can be used to maximize the likelihood $L_C(\boldsymbol{\theta}, \boldsymbol{\gamma})$. The resultant MLE of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ preserves the desired asymptotic properties.

We have developed a stand-alone computer program that implements the new methods. The program is reasonably efficient in terms of computation. It takes about 0.8 seconds to analyze one data set in the simulation studies presented in the next section on an iMac with a 3 GHz Intel Core i5 processor. This program is freely available on the website: https://sites.google.com/view/guoqingdiao-homepage.

Acknowledgements
We thank the referees for very helpful comments that improved the presentation of this article.

Authors' contributions
GD and DL participated in the design of the study. GD implemented the methods and drafted the manuscript. DL participated in the conception of the study and in editing the manuscript. Both authors read and approved the final manuscripts.

Funding
Not applicable.

Availability of data and materials
Data used in this article comes from the Childhood Asthma Management Program Ancillary Genetic Study.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Biostatistics and Bioinformatics, The George Washington University, Washington, District of Columbia, USA. [2] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

Received: 10 September 2019 Accepted: 17 August 2020
Published online: 07 September 2020

References

1. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437(7063):1299–320.
2. Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain?. Eur J Hum Genet. 2001;9(4):291.
3. Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. Genetic analysis of case/control data using estimated haplotype frequencies: application to apoe locus variation and alzheimer's disease. Genome Res. 2001;11(1):143–51.
4. Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol. 2002;23(3):221–33.
5. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33(3s):228.
6. Schaid DJ. Evaluating associations of haplotypes with traits. Genet Epidemiol. 2004;27(4):348–64.
7. Clark AG. Inference of haplotypes from pcr-amplified samples of diploid populations. Mol Biol Evol. 1990;7(2):111–22.
8. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. 1995;12(5): 921–7.
9. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 2001;68(4):978–89.
10. Zhang S, Pakstis AJ, Kidd KK, Zhao H. Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet. 2001;69(4):906–12.
11. Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet. 2002;70(1): 157–69.
12. Qin ZS, Niu T, Liu JS. Partition-ligation–expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet. 2002;71(5):1242–7.
13. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet. 2002;70(2):425–34.
14. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered. 2002;53(2):79–91.
15. Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet. 2003;73(6): 1316–29.
16. Lake SL, Lyon H, Tantisira K, Silverman E, Weiss S, Laird N, Schaid DJ. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered. 2003;55(1):56–65.

17. Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. Modeling and em estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered. 2003;55(4):179–90.
18. Lin D. Haplotype-based association analysis in cohort studies of unrelated individuals. Genet Epidemiol. 2004;26(4):255–64.
19. Lin D, Zeng D, Millikan R. Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet Epidemiol. 2005;29(4):299–312.
20. Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol. 2005;29(2):108–27.
21. Lin D, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies. J Am Stat Assoc. 2006;101(473):89–104.
22. Liu N, Beerman I, Lifton R, Zhao H. Haplotype analysis in the presence of informatively missing genotype data. Genet Epidemiol. 2006;30(4): 290–300.
23. Zhang K, Sun F, Zhao H. Haplore: a program for haplotype reconstruction in general pedigrees without recombination. Bioinformatics. 2004;21(1):90–103.
24. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet. 1996;58(6):1347.
25. Li J, Jiang T. Efficient rule-based haplotyping algorithms for pedigree data. In: Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB03). New York: ACM; 2003. p. 197–206.
26. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2001;30(1):97.
27. Zhang K, Zhao H. A comparison of several methods for haplotype frequency estimation and haplotype reconstruction for tightly linked markers from general pedigrees. Genet Epidemiol. 2006;30(5):423–37.
28. Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM. Family-based tests for associating haplotypes with general phenotype data: Application to asthma genetics. Genet Epidemiol. 2004;26(1):61–69.
29. 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(4):211–23.
30. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B (Methodol). 1977;39(1):1–22.
31. Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered. 2008;66(2):87–98.
32. Chen W-M, Manichaikul A, Rich SS. A generalized family-based association test for dichotomous traits. Am J Hum Genet. 2009;85(3): 364–76.
33. Lasky-Su J, Murphy A, McQueen MB, Weiss S, Lange C. An omnibus test for family-based association studies with multiple snps and multiple phenotypes. Eur J Hum Genet. 2010;18(6):720.
34. Huang Y-H, Lee M-H, Chen WJ, Hsiao CK. Using an uncertainty-coding matrix in bayesian regression models for haplotype-specific risk detection in family association studies. PLoS ONE. 2011;6(7):21890.
35. Lin D, Hu Y, Huang B. Simple and efficient analysis of disease association with missing genotype data. Am J Hum Genet. 2008;82(2):444–52.
36. Burdick JT, Chen W-M, Abecasis GR, Cheung VG. In silico method for inferring genotypes in pedigrees. Nat Genet. 2006;38(9):1002.
37. Chen W-M, Abecasis GR. Family-based association tests for genomewide association scans. Am J Hum Genet. 2007;81(5):913–26.
38. Group CAMPR, et al. The childhood asthma management program (camp): design, rationale, and methods. Control Clin Trials. 1999;20(1): 91–120.
39. Tashkin DP, Conolly ME, Deutsch RI, Hui KK, Littner M, Scarpace P, Abrass I. Subsensitization of beta-adrenoceptors in airways and lymphocytes of healthy and asthmatic subjects. Am Rev Respir Dis. 1982;125(2):185–93.
40. Silverman EK, Kwiatkowski DJ, Sylvia JS, Lazarus R, Drazen JM, Lange C, Laird NM, Weiss ST. Family-based association analysis of β2-adrenergic receptor polymorphisms in the childhood asthma management program. J Allergy Clin Immunol. 2003;112(5):870–6.
41. Abecasis GR, Cardon LR, Cookson W. A general test of association for quantitative traits in nuclear families. Am J Hum Genet. 2000;66(1):279–92.
42. Abecasis GR, Cookson WO, Cardon LR. Pedigree tests of transmission disequilibrium. Eur J Hum Genet. 2000;8(7):545.
43. Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet. 2005;76(4):592–608.
44. Diao G, Lin D. A powerful and robust method for mapping quantitative trait loci in general pedigrees. Am J Hum Genet. 2005;77(1):97–111.
45. Huang B, Amos C, Lin D. Detecting haplotype effects in genomewide association studies. Genet Epidemiol Off Publ Int Genet Epidemiol Soc. 2007;31(8):803–12.
46. Little RJA, Rubin DB. Statistical Analysis withMissing Data, 2nd Ed. Hoboken: Wiley; 2002.
47. Weir S. Genetic data analysis II. Sunderland, Massachusetts: Sinauer Associates; 1996, pp. 161–73.
48. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numeric Recipes in C: The Art of Scientific Computing, 2nd Ed. New York: Cambridge: Cambridge University Press; 1992.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.