Annals of
human genetics

doi: 10.1111/j.1469-1809.2010.00593.x

Haplotype Misclassification Resulting from Statistical Reconstruction and Genotype Error, and Its Impact on Association Estimates

Claudia Lamina[1,2], Helmut Küchenhoff[3], Jenny Chang-Claude[4], Bernhard Paulweber[5], H.-Erich Wichmann[1,6,7], Thomas Illig[1], Margret R. Hoehe[8], Florian Kronenberg[2] and Iris M. Heid[1,9]*

[1] *Institute of Epidemiology, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany*
[2] *Division of Genetic Epidemiology; Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria*
[3] *Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany*
[4] *Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[5] *First Department of Internal Medicine, Paracelsus Private Medical University Salzburg, Austria*
[6] *Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Munich, Germany*
[7] *Klinikum Grosshadern, Munich, Germany*
[8] *Max Planck Institute for Molecular Genetics, Berlin, Germany*
[9] *Department of Epidemiology and Preventive Medicine, Regensburg University Medical Center, Germany*

Summary

Haplotypes are an important concept for genetic association studies, but involve uncertainty due to statistical reconstruction from single nucleotide polymorphism (SNP) genotypes and genotype error. We developed a re-sampling approach to quantify haplotype misclassification probabilities and implemented the MC-SIMEX approach to tackle this as a 3 × 3 misclassification problem. Using a previously published approach as a benchmark for comparison, we evaluated the performance of our approach by simulations and exemplified it on real data from 15 SNPs of the *APM1* gene. Misclassification due to reconstruction error was small for most, but notable for some, especially rarer haplotypes. Genotype error added misclassification to all haplotypes resulting in a non-negligible drop in sensitivity. In our real data example, the bias of association estimates due to reconstruction error alone reached −48.2% for a 1% genotype error, indicating that haplotype misclassification should not be ignored if high genotype error can be expected. Our 3 × 3 misclassification view of haplotype error adds a novel perspective to currently used methods based on genotype intensities and expected number of haplotype copies. Our findings give a sense of the impact of haplotype error under realistic scenarios and underscore the importance of high-quality genotyping, in which case the bias in haplotype association estimates is negligible.

Keywords: Haplotypes, measurement error, genotyping error, association studies, misclassification

Introduction

Haplotype association studies have gained influence due to the availability of high-density single nucleotide polymorphism (SNP) data and their strengths in multi-locus analysis. Particularly, with the availability of genome-wide SNP genotype data, an increase in the number of haplotype association studies utilizing this large-scale genotype data for haplotype association studies can be anticipated, following the current exploitation of these data in genotype association studies.

A haplotype covers a DNA-sequence on one chromosome that is often inherited jointly. Haplotypes contain epistatic information on multiple SNPs, are expected to contain information on an intervening un-genotyped causal variant, and may represent the more biologically relevant entity (Clark,

*Corresponding author: Iris. M. Heid, Department of Epidemiology and Preventive Medicine, Regensburg University Medical Center, Franz-Josef-Strauß-Allee 11, 93053 Regensburg, Germany. Tel: +49 941 9445210; Fax: +49 941 9445252; E-mail: iris.heid@klinik.uni-regensburg.de

2004; Schaid, 2004). Haplotypes also reduce data complexity, as the number of haplotypes appearing in a population usually undercut the number of theoretically possible haplotypes (Daly et al., 2001; Johnson et al., 2001) and thus provide a power gain (Akey et al., 2001; Morris & Kaplan, 2002). Since determining haplotypes in the laboratory is still not practicable for large epidemiological studies, haplotypes are usually reconstructed statistically from SNP genotypes, e.g., via the EM-algorithm (Excoffier & Slatkin, 1995) involving estimation uncertainty (Lamina et al., 2008).

Genotype error adds to this uncertainty: The error from the genotyping process itself ("*genotyping error*") is just one of many sources of error (Pompanon et al., 2005). The pure "genotyping error" can be estimated from repeated genotyping and is reportedly small for established genotyping methods such as multiplex approaches (0.01%–1% (Ranade et al., 2001), 0.1% (Heid et al., 2008)) but might be larger for more recently established genotyping methods such as genome-wide SNP chip-genotyping. While the impact of such a genotyping error is shown to be negligible for single SNP association (Heid et al., 2008), it is an open question to what extent the genotype error accumulates across multiple loci resulting in more substantial haplotype error.

Haplotype error can result in misclassifying subjects: When subjects are chosen according to their assigned haplotypes for in-depth and expensive functional studies, misclassified haplotypes misclassify subjects and compromise functional studies. Haplotype error can also result in biased haplotype association estimates: It is well known that errors in explanatory variables induce bias in regression estimates and reduce power (Carroll et al., 2006). Non-negligible bias in haplotype association estimates was reported in simulation studies by pure reconstruction error assuming perfect genotypes (Kraft et al., 2005) and by pure genotype error assuming unambiguous reconstruction and a genotype error of 5% (Govindarajulu et al., 2006). Thus, haplotype uncertainty due to both reconstruction and genotype error under realistic scenarios and its impact on haplotype assignment and on association estimates is not well understood, particularly for real data situations.

Most previously reported approaches to account for haplotype error use the estimated number of copies of haplotypes in the association analyses (Lake et al., 2003; Schaid 2004), which transforms the biologically trichotomous haplotype variable (zero, one, or two copies of a haplotype) into a continuous entity. Here, we view haplotype error as a 3 × 3 misclassification problem defined by the misclassification probabilities: We present a re-sampling approach to estimate the misclassification probabilities. We implement the "*misclassification simulation and extrapolation*" (MC-SIMEX), an approach to account for misclassification in a general model framework allowing for covariate adjustment and for a wide variety of misclassification schemes and association analysis models (Kuchenhoff et al., 2006). We test this approach via simulation studies and apply it to a real data example of 15 SNPs of the *APM1* gene from 1770 subjects with plasma adiponectin concentrations of the SAPHIR study (Heid et al., 2006). In both simulations as well as in our real data example, we compare our approach with one of the most widely used methods for haplotype association analysis accounting for haplotype error (Lake et al., 2003).

It was the aim of our investigation to estimate misclassification probabilities of haplotype error from reconstruction and genotype error, to present sensitivity and specificity for haplotype assignment, and the bias in haplotype association estimates under realistic scenarios, and to provide an approach to tackle this as a 3 × 3 misclassification problem.

Methods

The 3 × 3 Misclassification Problem

Let $G_i = (G_{i1}, \ldots, G_{iL})$ denote the *true genotype* for the *i*th subject, $i = 1, \ldots, n$, for L SNPs with G_{il} indicating the number of minor alleles at locus l, $l = 1, \ldots, L$, for individual i ($G_{il} \in \{0, 1, 2\}$) and the genotype probabilities $\pi_G^{(l)} = (\pi_{G,0}^{(l)}, \pi_{G,1}^{(l)}, \pi_{G,2}^{(l)})$. Accordingly, $G_i^* = (G_{i1}^*, \ldots, G_{iL}^*)$ denote the *observed error-prone genotypes* with genotype probabilities $\pi_G^{(l)*} = (\pi_{G,0}^{(l)*}, \pi_{G,1}^{(l)*}, \pi_{G,2}^{(l)*})$, which are estimated by the observed genotype frequencies. Assuming an allele-independent genotype error implies that the probability of misclassifying the major allele A as the minor allele a equals the probability of misclassifying the minor allele as the major, $P(A \to a) = P(a \to A) = \varepsilon$.

For $M = 2^L$ haplotypes $h_1, \ldots, h_M$, the haplotypes of subject i can be written as $H_i = (H_{i1}, \ldots, H_{iM})$, with H_{im} indicating the *true number of copies of haplotype* h_m *of subject i*, $m = 1, \ldots, M$ ($H_{im} \in \{0, 1, 2\}$). Accordingly, the (observed) *reconstructed number of copies of each haplotype of subject i* is $H_i^* = (H_{i1}^*, \ldots, H_{iM}^*)$, which is derived by the expected values given the observed genotypes, $E(H|G^*)$, for example, using the EM algorithm ($H_{im}^* \in [0, 2]$); the (observed*) most likely number of haplotypes* $C_i^* = (C_{i1}^*, \ldots, C_{iM}^*)$ *of subject i* is derived by categorizing H_i^* into the most likely haplotype pair for each individual, thus returning to the discrete space ($C_{im}^* \in \{0, 1, 2\}$). For each haplotype h_m, the transition $H_{im} \to C_{im}^*$ describes a classical 3 × 3 misclassification problem denoted by the *misclassification probabilities* $\pi_{kl}^{(m)} = P(C_{im}^* = k | H_{im} = l)$, $k, l = 0, 1, 2$, which form the matrix $\Pi^{(m)} = (\pi_{kl}^{(m)})_{k,l=0,1,2}$ involving six unknown parameters due to $\pi_{0l}^{(m)} + \pi_{1l}^{(m)} + \pi_{2l}^{(m)} = 1$ for $l = 0, 1, 2$.

In the case of no genotype error, the subjects truly having two copies of the same haplotype ("*true homozygous*") can always be reconstructed correctly from the genotypes, as the genotypes are then homozygous at all loci, and thus $\pi_{02}^{(m)}$ and $\pi_{12}^{(m)}$ equal zero. Also, when the reconstructed haplotype pair for a subject involves two copies of the same haplotype ("*observed homozygous*"), this implies homozygous genotypes at all loci und thus unambiguous reconstruction. Hence $\pi_{20}^{(m)}$ and

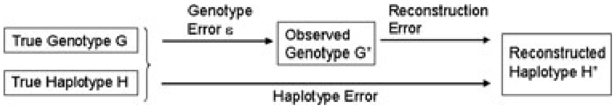

Figure 1 Schematic overview of the haplotype error sources.

$\pi_{21}^{(m)}$ equal zero. The misclassification problem therefore reduces to two unknown parameters $\pi_{00}^{(m)}$ and $\pi_{01}^{(m)}$. This can be re-parameterized by the sensitivity $Sn_m = P(C_{im}^* > 0|H_{im} > 0)$ ("the probability that a true carrier of a certain haplotype is classified correctly") and the specificity $Sp_m = P(C_{im}^* = 0|H_{im} = 0)$ ("the probability that a true non-carrier is classified correctly") via $\pi_{00}^{(m)} = Sp_m$ and $\pi_{01}^{(m)} = \frac{\pi_1^{(m)}+\pi_2^{(m)}-Sn_m(\pi_1^{(m)}+\pi_2^{(m)})}{\pi_1^{(m)}}$. Here, $\pi_k^{(m)} = P(H_{im} = k)$denotes the true probability for a subject having k copies of one haplotype h_m.

When genotype error $G_i \to G_i^*$ is involved in the misclassification problem $H_{im} \to C_{im}^*$ (see Fig. 1), haplotypes that would have been unambiguous through reconstruction alone are now also subject to error and $\pi_{02}^{(m)}$, $\pi_{12}^{(m)}$, $\pi_{20}^{(m)}$, and $\pi_{21}^{(m)}$ may deviate from zero leaving six parameters for misclassification probability estimation. The sensitivity and specificity can be determined from the misclassification probabilities as $Sn_m = \frac{\pi_{11}^{(m)}\pi_1^{(m)}+\pi_{12}^{(m)}\pi_2^{(m)}+\pi_{21}^{(m)}\pi_1^{(m)}+\pi_{22}^{(m)}\pi_2^{(m)}}{\pi_1^{(m)}+\pi_2^{(m)}}$ and $Sp_m = \pi_{00}^{(m)}$. If a dominant genetic effect is assumed, the 3 × 3-misclassification matrix reduces to a 2 × 2 problem and is then, again, completely determined by sensitivity and specificity, even if genotype error is involved.

While sensitivity and specificity are haplotype-specific error measures, the *overall error* (i.e., the proportion of subjects with falsely classified haplotypes) summarizes overall haplotypes, $ER_{all} = \sum_{i=1}^{n}(1 - c_i)/N$ where $c_i = 1$, if $H_i = C_i^*$, otherwise $c_i = 0$. The *overall discrepancy* D depicts the error in haplotype frequencies. It is defined as the proportion of differences between observed haplotype frequencies $(\hat{f}_1^*, \ldots, \hat{f}_M^*)$ and true haplotype frequencies $(\hat{f}_1, \ldots, \hat{f}_M)$: $D = D(\hat{f}_1, \ldots, \hat{f}_M, \hat{f}_1^*, \ldots, \hat{f}_M^*) = \frac{1}{2}\sum_{m=1}^{M}|\hat{f}_m - \hat{f}_m^*|$, ranging between 0 and 1 (Stephens et al., 2001). The discrepancy can also be depicted in a haplotype-specific way: $D_m(\hat{f}_m, \hat{f}_m^*) = \frac{1}{2}|\hat{f}_m - \hat{f}_m^*|$.

Estimating the Misclassification Probabilities via Re-sampling

An approach to derive haplotype misclassification probabilities in its most general form for given SNP genotype data was developed: As haplotype-specific discrepancies were shown to be negligibly small (Lamina et al., 2008), we assumed that observed haplotype frequencies reasonably approximated true haplotype frequencies. For each simulation run, 1000 haplotypes were randomly drawn given this true haplotype probability distribution $f = (f_1, \ldots, f_M)$. Two haplotypes were randomly assigned to each of 500 subjects assuming random mating. From each subject's true haplotypes, the subjects' genotypes were deduced. The genotypes were then subjected to genotype error with $\varepsilon = 0\%$, 0.5%, or 1% for each allele, which implies that a subject with a true homozygous genotype at one SNP is assigned a heterozygous or the other homozygous genotype with probability $2\varepsilon(1 - \varepsilon)$ or ε^2, respectively. From these error-prone genotypes, haplotypes were reconstructed using the EM algorithm "proc haplotype" (SAS, Heidelberg, Germany) and compared with the true haplotypes using the error measures sensitivity, specificity, misclassification probabilities, and discrepancies. For 100 simulations, the mean and the standard deviation of these error measures were computed.

Evaluating the Performance of the MC SIMEX in Haplotype Association Analyses

We estimated the misclassification matrix Π using the re-sampling approach as described above based on the real data haplotype frequencies. In the MC-SIMEX approach, data are simulated with increasing misclassification $\Pi^{1+\lambda}$, $\lambda = 1, 2, \ldots$, and association estimates $\hat{\beta}_\lambda^*$ are computed starting with the observed data set ($\lambda = 0$) and the observed association estimate $\hat{\beta}^* = \hat{\beta}_0$ (simulation step). Then, a function (linear, quadratic, or log-linear) is fitted to the β-estimates and extrapolated back to the case of no misclassification for $\lambda = -1$ (extrapolation step), which is the SIMEX-corrected estimate (see Fig. S1 for illustration). The MC-SIMEX can be applied to all Generalized Linear Models (GLM) for basically any given misclassification matrix and is implemented in R (package "simex").

We simulated normally distributed outcome data for linear regression analysis for 1000 subjects: We assumed a risk haplotype of interest, h_R, with population probability f_{h_R} and denoted any other haplotype by h. Subjects were thus assigned the h/h, h_R/h, or h_R/h_R haplotype pair with probabilities $(1 - f_{h_R})^2$, $2f_{h_R}(1 - f_{h_R})$, or $f_{h_R}^2$ and their outcome values were drawn from $N(0, \sigma^2)$, $N(\beta, \sigma^2)$, or $N(2\beta, \sigma^2)$, respectively, assuming additivity of the effect β per copy of h_R,. The variance of the outcome, σ^2, was set to 0.4 to mimic adiponectin plasma level (on the log(adiponectin+1) scale) from the real data example. Different scenarios included effect estimates β of 0.5 or 0.05, haplotype probabilities f_{h_R} of 0.15 and 0.3, and two misclassification schemes,

$$\Pi_{low} = \begin{pmatrix} 0.975 & 0.1 & 0.01 \\ 0.025 & 0.9 & 0.1 \\ 0 & 0 & 0.89 \end{pmatrix} \quad \text{and}$$

$$\Pi_{high} = \begin{pmatrix} 0.899 & 0.3 & 0.1 \\ 0.1 & 0.69 & 0.3 \\ 0.001 & 0.01 & 0.6 \end{pmatrix}.$$

Based on the assigned true haplotype pairs h_R/h_R, h_R/h, or h/h for each subject and applying the haplotype misclassification schemes Π_{low} or Π_{high},, the observed haplotype pairs h_R^*/h_R^*, h_R^*/h or h^*/h were obtained. To evaluate the performance of the MC-SIMEX in correcting for haplotype misclassification and to check the preservation of the additivity of the effect, the β-estimates comparing the subjects with h_R^*/h^* ($\hat{\beta}_1$) or h_R^*/h_R^* ($\hat{\beta}_2$) with h^*/h^* subjects were computed (i) ignoring the haplotype misclassification, (ii) accounting for it by the MC-SIMEX approach, and (iii) accounting for it by our benchmark method (Lake et al., 2003) as implemented in R (haplo.glm). In the method by Lake and colleagues, haplotypes and haplotype association are estimated in a single step incorporating the outcome variable and covariate information.

For the 200 simulation runs, we computed mean and standard deviation of effect estimates, 95% coverage (i.e., the proportion of 95% confidence intervals that contain the true effect), and relative bias, $((\hat{\beta}_{naive} - \beta)/\beta)$*100%, with $\hat{\beta}_{naive}$ being the naive estimator.

Real Data Example

To provide a real data example, we re-analyzed data from the SAPHIR study (Salzburg Atherosclerosis Prevention Program in Subjects at High Individual Risk), an observational study involving 1770 healthy unrelated subjects. From the known 53 SNPs of the adiponectin encoding *APM1* gene genotyped in a subsample, 15 haplotype tagging SNPs with minor allele frequencies of >1% were selected according to Stram et al. (2003) and genotyped in the full sample. Haplotypes were reconstructed via the EM algorithm (SAS proc haplotype). See Heid et al. (2006) for details and notation of haplotypes.

We derived approximate misclassification matrices for each haplotype combining the reconstruction error and a genotype error of 0% (pure reconstruction error), 0.5% and 1%. We computed haplotype linear regression estimates (i) ignoring the haplotype misclassification using the naive estimator based on C*, (ii) accounting for it by the MC-SIMEX, or (iii) accounting for it by the method from Lake and colleagues (Lake et al., 2003). The linear model was computed on log(adiponectin+1) adjusted for age, sex, body mass index (BMI), and all other haplotypes with the most frequent one as reference (H22, frequency = 0.124).

Results

Estimating Misclassification Probabilities via Re-sampling

Exemplified on the *APM1* data, we derived the haplotype misclassification probabilities given the observed genotype data, the observed haplotype frequencies, and the assumed genotype error via re-sampling. In this data, 18 of the 43 reconstructed *APM1* haplotypes had frequencies >1%. Table 1 depicts the haplotype misclassification matrices with and without genotype error for selected haplotypes (all haplotype misclassification matrices are shown in Table S1). As expected, π_{02}, π_{12}, π_{20}, and π_{21} were zero in the case of no genotype error. When adding a genotype error of 0.5%

Table 1 Misclassification matrices for selected *APM1* haplotypes assuming 0, 0.5, and 1% genotype error per allele (more details are given in Table S1).

Haplotype (frequency)	Genotype error														
	0%					0.5%					1%				
H16 (0.100)			H					H					H		
			0	1	2			0	1	2			0	1	2
	C*	0	0.9726	0.0436	0	C*	0	0.9642	0.1141	0	C*	0	0.9589	0.1754	0.0058
		1	0.0274	0.9564	0		1	0.0358	0.8857	0.1438		1	0.0411	0.824	0.2406
		2	0	0	1		2	0	0.0002	0.8562		2	0	0.0006	0.7535
H2 (0.053)			H					H					H		
			0	1	2			0	1	2			0	1	2
	C*	0	0.9947	0.0070	0	C*	0	0.9944	0.0852	0	C*	0	0.9941	0.1672	0
		1	0.0053	0.9930	0		1	0.0056	0.9148	0.0934		1	0.0059	0.8328	0.1833
		2	0	0	1		2	0	0	0.9066		2	0	0	0.8167
H12 (0.023)			H					H					H		
			0	1	2			0	1	2			0	1	2
	C*	0	0.9947	0.1030	0	C*	0	0.9936	0.2096	0	C*	0	0.9929	0.2902	0.0345
		1	0.0053	0.8970	0		1	0.0064	0.79	0.2586		1	0.0071	0.7091	0.3276
		2	0	0	1		2	0	0.0004	0.7414		2	0	0.0007	0.6379
H4 (0.019)			H					H					H		
			0	1	2			0	1	2			0	1	2
	C*	0	1	0.0185	0	C*	0	0.9976	0.1363	0	C*	0	0.9955	0.2265	0
		1	0	0.9815	0		1	0.0024	0.8637	0		1	0.0045	0.7735	0
		2	0	0	1		2	0	0	1		2	0	0	1

or 1%, all misclassification probabilities (for $i \neq j$) increased and most of the π_{02}, π_{12}, π_{20}, and π_{21} deviated from zero. For example, for the common haplotype H16, the misclassification probability from pure reconstruction error was rather moderate with up to 4.36%, while misclassification increased up to 24.06% when adding genotype error. The extent of the haplotype reconstruction error was more substantial for some (e.g., H12 with up to 10.30%), but not all (e.g., H4 with up to 1.85%) rarer haplotypes, which was already noted previously (Lamina et al., 2008). Genotype error added markedly to all haplotypes. Overall, the genotype error contributed more substantially to the overall misclassification than the pure reconstruction error: The overall misclassification error of all reconstructed haplotypes increased from 6.67% in the case of no genotype error to 20.15% or 31.36% in the case of 0.5% or 1% genotype error, respectively.

Summarizing the misclassification by sensitivity and specificity illustrated the dependence of pure reconstruction error on haplotype frequencies: The sensitivity was high for the common haplotypes and for many rare haplotypes, but was substantially decreased down to 50% for some rare haplotypes (Fig. 2A, Table S2). It can further be seen that the genotype error reduced the sensitivity independently from haplotype frequency. The sensitivity was as low as 40% for some haplotypes with a genotype error of 1%. The specificity was reduced for the common haplotypes, but was 100% for most rare haplotypes and decreased in the presence of genotype error but never fell below 95% (Fig. 2B).

Haplotype-specific discrepancies were small: they did not exceed 0.001 in the case of no genotype error and reached a maximum at 0.006 for 1% genotype error (Fig. 3). Overall discrepancy increased for increasing genotype error of 0.5% or 1% from 0.0199 (no genotype error) to 0.0738 or 0.1253, respectively.

It should be noted that haplotype reconstruction error and genotype error not only evoke misclassified haplotypes, which is grasped by the misclassification matrix, but also gave rise to "newly created" haplotypes: The percentage of falsely created haplotypes increased from 0.395% for no genotype error to 6.30% or 11.50% for 0.5% or 1% genotype error, respectively. However, the frequencies of falsely created haplotypes did not exceed 0.25% and these haplotypes would usually not enter haplotype association analyses due to sparseness of data.

Simulation Studies: Bias in Estimates and Performance of MC-SIMEX

Simulation results to judge the performance of the MC-SIMEX compared to the true, the naive, or the haplo.glm model by Lake and colleagues are summarized in Table 2. We have also compared the performance of the various SIMEX extrapolation functions and found the loglinear function to perform best (Table S3). While the estimates of the naive model clearly underestimated the true haplotype effects, the SIMEX estimates approximated the true estimate very well. Estimators from the haplo.glm model still underestimated the true association almost to the same extent as the naive method, which can be explained by the fact that the largest contribution to the misclassification was from the genotype error (as described above), which is neglected by the haplo.glm approach.

The relative bias for β_1 ranged between −8.9% and −10.2% (Π_{low}) or −31.4% and −36.9% (Π_{high}) and for β_2, between −1.9% and −4.2% (Π_{low}) or −17.3% and −22.1% (Π_{high}). We observed no dependence of the bias on risk haplotype frequency. Coverage of 95% confidence intervals across the 200 simulations indicated that type I error was well preserved.

APM1 Real Data: Correcting Haplotype Association Estimates for Misclassification

Figure 4A shows naive and corrected beta-estimates for the three most common haplotypes (frequencies > 10%, except the most common haplotype serving as reference) of the *APM1* real data example. Estimates increased when correcting for pure reconstruction error by the MC-SIMEX and further increased when additionally accounting for genotype error. The correction using the haplo.glm model by Lake and colleagues yielded similar beta-estimates as the SIMEX-correction for pure reconstruction error, which is as expected as the haplo.glm model does not incorporate the genotype error. For example, for the haplotype H16, the β_1 estimate was 0.086 without correction (β_2: 0.224), 0.104 with pure reconstruction MC-SIMEX correction (β_2: 0.239), 0.118 correcting for 0.5% genotyping error with MC-SIMEX (β_2: 0.273), 0.130 correcting for 1% genotyping error with MC-SIMEX (β_2: 0.297), and 0.095 (β_2: 0.253) with haplo.glm. Thus, haplo.glm estimates are comparable with the estimates corrected for pure reconstruction error. The relative bias ranged from −3.1 to −20.7% (mean −10.2%) when correcting for pure reconstruction error, while it ranged from −31.0 to −38.9% (mean −16.2%) or −54.5% to −48.2% (mean −20.6%) when adding 0.5% or 1% genotype error, respectively. It should be noted that additivity of the genetic effect did not fully hold.

A similar picture can be seen for the four haplotypes with modest frequency (between 5% and 10%). For haplotypes with frequencies <5%, however, the picture was less consistent. Since covariate information is additionally used for haplotype reconstruction, this could be due to the limited number of subjects available for such haplotypes together with specific covariate combinations (Fig. 4B).

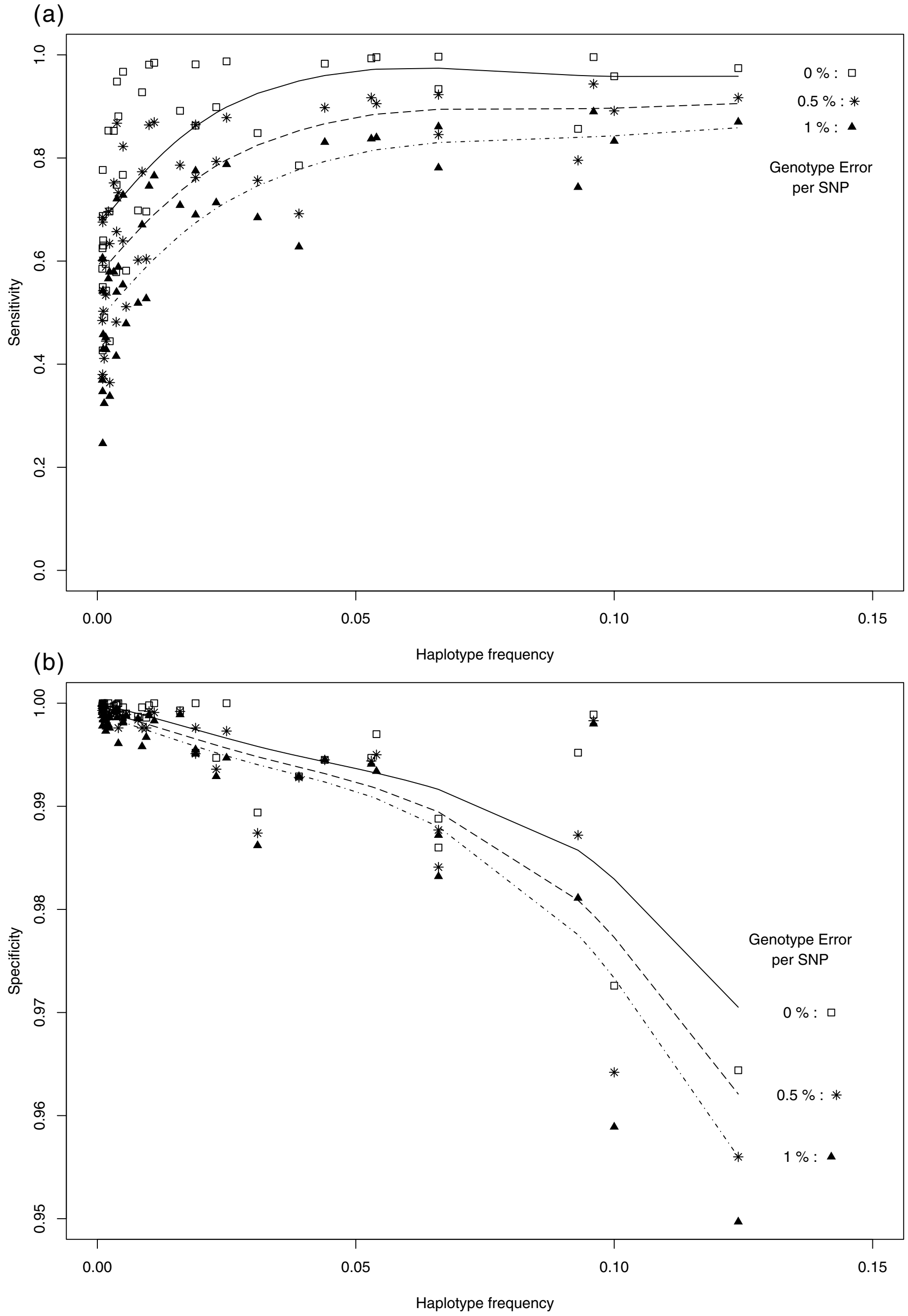

Figure 2 (A) Sensitivity and (B) Specificity as a function of haplotype frequencies in *APM1* gene haplotypes for varying genotype errors (0, 0.5, and 1%).

Discussion

We introduce haplotype error as a 3 × 3 misclassification problem and provide a unified approach to account for this misclassification in haplotype assignment and haplotype association analyses. We provide a re-sampling approach to estimate misclassification, present sensitivity, and specificity as measures of classification error in haplotype assignment, and introduce the MC-SIMEX approach to account for haplotype error in association analyses. Our approach allows general modeling and shows similar performance as the most widely used approach by Lake et al. (2003) to correct for pure

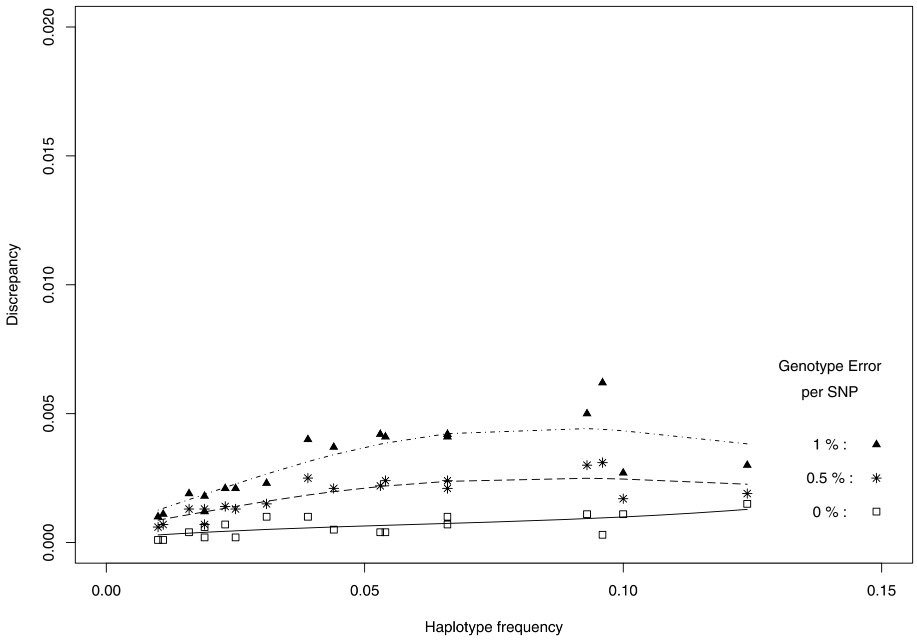

Figure 3 Discrepancy D_m as a function of haplotype frequencies in APM1 gene haplotypes for varying genotype errors (0, 0.5, and 1%).

Table 2 Performance of MC-SIMEX correction of haplotype misclassification in simulation studies: A risk haplotype h_R was simulated and tested for association with a quantitative phenotype using a 2df model (estimating β_1, β_2) for various haplotype probabilities f_{h_R} and misclassification schemes Π_{low} or Π_{high}. Across 200 simulated data sets, mean effect and coverage of 95% confidence intervals are given for the true model, the model ignoring haplotype error (naive model), accounting for the error as proposed by Lake et al. (2003) (haplo.glm model) and accounting for the error using the SIMEX corrected estimates using the loglinear function.

Model	Misclassification matrix Π_{low}		Misclassification matrix Π_{high}	
	Mean($\hat{\beta}_1$) [Coverage]	Mean($\hat{\beta}_2$) [Coverage]	Mean($\hat{\beta}_1$) [Coverage]	Mean($\hat{\beta}_2$) [Coverage]
$\beta_1 = 0.5$, $\beta_2 = 1$, $\sigma^2 = 0.4$, $f_{h_R} = 0.15$				
True	0.5028 [0.94]	1.0031 [0.98]	0.5044 [0.97]	1.0017 [0.94]
Naive	0.4536 [0.63]	0.9844 [0.99]	0.3183 [0.00]	0.8238 [0.67]
Haplo.glm	0.4565 [0.69]	0.9847 [0.99]	0.3209 [0.00]	0.8241 [0.67]
Simex (loglin)	0.5041 [0.94]	0.9998 [0.98]	0.5003 [0.83]	1.0922 [0.81]
$\beta_1 = 0.5$, $\beta_2 = 1$, $\sigma^2 = 0.4$, $f_{h_R} = 0.3$				
True	0.5018 [0.95]	0.9959 [0.96]	0.4999 [0.93]	0.9974 [0.96]
Naive	0.4547 [0.63]	0.9538 [0.85]	0.3429 [0.00]	0.8244 [0.20]
Haplo.glm	0.4555 [0.66]	0.9539 [0.86]	0.3435 [0.00]	0.8245 [0.20]
Simex (loglin)	0.4963 [0.94]	0.9854 [0.93]	0.4881 [0.82]	0.9741 [0.94]
$\beta_1 = 0.05$, $\beta_2 = 0.10$, $\sigma^2 = 0.4$, $f_{h_R} = 0.15$				
True	0.0528 [0.94]	0.1031 [0.98]	0.0544 [0.97]	0.1017 [0.94]
Naive	0.0481 [0.96]	0.1003 [0.97]	0.0346 [0.91]	0.0817 [0.91]
Haplo.glm	0.0484 [0.96]	0.1004 [0.97]	0.0349 [0.91]	0.0818 [0.91]
Simex (loglin)	0.0537 [0.95]	0.1029 [0.98]	0.0557 [0.85]	0.1108 [0.87]
$\beta_1 = 0.05$, $\beta_2 = 0.10$, $\sigma^2 = 0.4$, $f_{h_R} = 0.3$				
True	0.0518 [0.95]	0.0959 [0.97]	0.0499 [0.93]	0.0974 [0.96]
Naive	0.0465 [0.95]	0.0918 [0.96]	0.0333 [0.92]	0.0759 [0.96]
Haplo.glm	0.0468 [0.93]	0.0919 [0.96]	0.0331 [0.92]	0.0758 [0.95]
Simex (loglin)	0.0509 [0.92]	0.0951 [0.95]	0.0502 [0.86]	0.0902 [0.91]

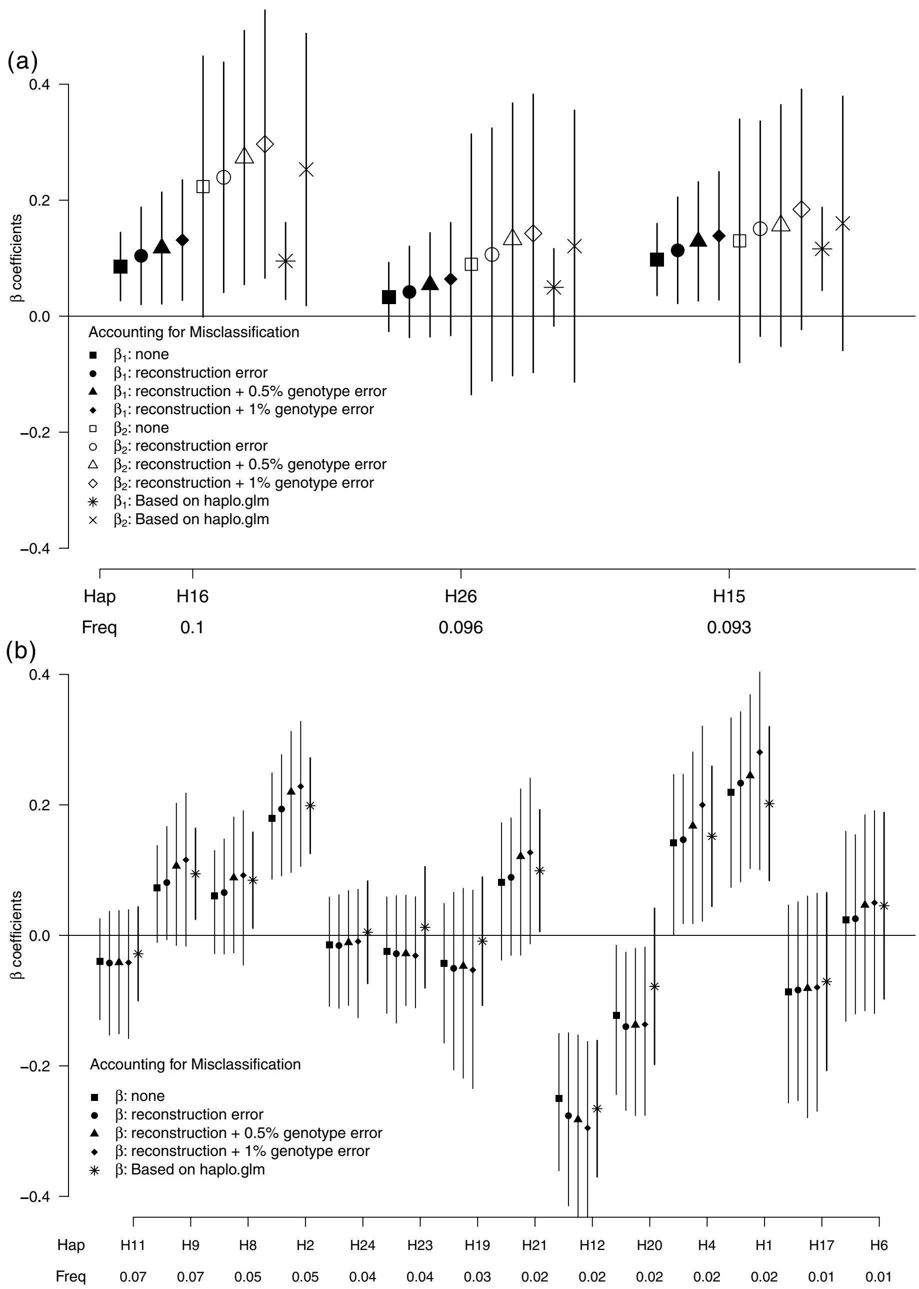

Figure 4 β-coefficients for 17 *APM1* gene haplotypes with frequency >1% compared to subjects with two copies of the most common haplotype H22 as reference, based on a linear regression model on log(adiponectin +1) adjusted for age, sex, body mass index, and all other haplotypes, without accounting for error (naive estimator), accounting for reconstruction error with 0%, 0.5%, and 1% genotype error using the SIMEX correction, and accounting for reconstruction error using the haplo.glm approach by Lake et al. (2003)) (A) for common haplotypes (frequencies > 10%) using a 2df genetic model, and (B) for rarer haplotypes (frequencies < 10%) collapsing the homozygote subjects of the rare allele– if any– with the heterozygote subjects.

reconstruction error. Our approach is at the same time flexible to additionally account for genotype error. We present both, simulation and real data results and quantify the haplotype misclassification under realistic scenarios.

Estimating Misclassification Under Realistic Scenarios

We found the pure reconstruction error to be small relative to the uncertainty added by a genotype error of 0.5% or 1%. This genotype error size is consistent with previous error estimates from double multiplex genotyping (Heid et al., 2008). It can be argued that more recently developed genotyping methods could have a higher genotype error, for which our estimates undercut the real bias. Interestingly, the haplotype uncertainty added through the genotype error appeared to be independent from haplotype frequencies in contrast to the haplotype reconstruction error (Lamina et al., 2008).

An alternative approach to assess haplotype misclassification probabilities might be molecular haplotyping (Levenstien et al., 2006). However, laboratory-assessed haplotypes are also subject to error possibly to a larger extent than SNP genotypes and cannot be considered a gold standard. Furthermore, laboratory-assessed haplotypes are too costly to be assessed in large scale, while our re-sampling approach to quantify haplotype misclassification can be performed without laboratory expenses in the routine epidemiological setting.

Sensitivity and Specificity as Measures of Classification Error for Haplotype Assignment

Sensitivity and specificity can be derived easily from the misclassification probabilities and provide an intuitive and well-accepted measure of classification error in life sciences. In our scenarios, sensitivity was down to 40%. If a researcher aims to select individuals with a certain haplotype for demanding functional studies, knowing the probability that a selected individual really has this haplotype can substantially guide the planning and success of such an investigation.

Performance of MC-SIMEX Correction Compared to Benchmark Method

The MC-SIMEX correction for pure reconstruction error compared well with the benchmark method by Lake et al., while at the same time providing the flexibility to incorporate genotype error. As expected, estimates accounting for the misclassification yielded estimates corrected "away from the null" compared to the naive estimates for almost all haplotypes and slightly extended confidence intervals.

Bias of Association Estimates Due to Haplotype Error

In our real data example, association estimates uncorrected for misclassification were underestimated by up to 50% assuming a genotype error of 1%. Our findings were thus in the same ballpark as in previous reports (Govindarajulu et al., 2006) describing an underestimation of up to 32.3% after accounting for 1% genotyping error, but ignoring haplotype reconstruction error. However, for most haplotypes and/or for high-quality genotypes ($\leq$0.5% error), the impact of haplotype misclassification on haplotype association estimates will be moderate and less relevant. This stresses the importance of high genotyping quality when estimating haplotypes, but also the validity of haplotype association analyses when based on good genotypes. It will be an important issue to quantify genotype error for genome-wide SNP chips' genotypes in order to understand the haplotype error, the underlying bias in haplotype association estimates, and the extent of the decreased power in genome-wide haplotype association analyses.

Usefulness of our 3 × 3 Misclassification Approach Compared to Existing Approaches

There are several available approaches to account for haplotype error: In the approach by Zhu and Guo (Zhu & Guo, 2006), haplotype reconstruction is based on fluorescent intensity genotype data instead of the called trichotomous genotype. However, in most cases in current practice, these genotype intensities will not be available to the analysts of epidemiological data. Methods for haplotype association analysis have been developed using H^* within the likelihood framework (Zaykin et al., 2002; Epstein & Satten, 2003; Lake et al., 2003; Spinka et al., 2005) or with estimating equations (Zhao et al., 2003). While the haplotype reconstruction error is accounted for in these analyses (Mensah et al., 2007), they do not incorporate genotype error.

Schaid et al. (2002) propose a score test using the expected number of copies, $E[H^*|G]$, according to the regression calibration idea. This approach usually does not cover genotype error, but could be extended. Lake et al. (2003) extended the approach by Schaid and colleagues by additionally including information on the outcome and covariates: Haplotypes are reconstructed in one step together with estimating β-coefficients. However, this "one-step" approach might not be preferable for all situations: If several outcomes are of interest, the haplotypes different for each outcome complicate comparisons across outcomes. If picking subjects with a specific haplotype combination is of interest for functional follow-up, the inclusion of the outcome in the haplotype estimation step might also not be ideal. We totally recognize that the approach by Lake and colleagues can be considered state of the art. Therefore, we included it as a "benchmark" method for comparing it with our approach.

While the MC-SIMEX method is an existing approach for generally accounting for misclassification (Kuchenhoff et al., 2006), it was never before applied to haplotype misclassification. In fact, accounting for haplotype error in combination with genotype error was never before viewed as a 3 × 3 misclassification problem. In combining this with a re-sampling approach to estimate misclassification probabilities taking into account both genotype error as well as statistical reconstruction error, we provide a unified approach to tackle haplotype misclassification. The 3 × 3 misclassification approach is quite natural and intuitive. It can add to the existing approaches depending on the research question for which the haplotypes are constructed: to provide measures of classification error for subject selection or whenever the specific numbers of copies of a haplotype or the specific haplotype pair are of interest.

Strengths and Limitations of This Investigation

It is a strength of our investigation that we applied our approach to high dense SNP genotype data with a complex underlying haplotype structure, which showed very high associations with a quantitative disease-relevant blood marker. Furthermore, we applied realistic scenarios to estimate haplotype misclassification.

It may be considered a limitation of our re-sampling approach that we assumed the observed haplotype frequencies to sufficiently approximate true haplotype frequencies. However, this seemed suitable, as haplotype-specific discrepancies were found to be very small throughout. Furthermore, we assumed the genotype error to be allele-independent (i.e., the probability of misclassifying the minor allele into the major allele being the same as the other way round). This implies a restricted genotype error model, which may not grasp real situations completely. However, it is the currently most widely applied genotype error model (Wong et al., 2004; Govindarajulu et al., 2006; Moskvina & Schmidt, 2006), and it has been shown to be reasonably applicable while being the most parsimonious model (Heid et al., 2008). In fact, our approach of the 3 × 3 haplotype misclassification has the advantage that it can be easily extended to incorporate a general genotype error model due to its compatibility with the 3 × 3 genotype misclassification problem.

Conclusions

Our investigation underscores that haplotype misclassification, as a result of genotype error and statistical reconstruction from these genotypes, can be substantial for some haplotypes and in the case of high genotype error, but also that bias from haplotype misclassification is small in the case of high-quality genotype data. We present the MC-SIMEX approach as an efficient method to correct association estimates for haplotype misclassification, which yields comparable results to the haplo.glm method by Lake et al. (2003), while providing full flexibility of models. Finally, we suggest that haplotype error may be a 3 × 3 misclassification problem in existing approaches, which can be of particular interest under specific research scenarios.

Acknowledgement

This work was supported by a grant of the "Deutsche Forschungsgesellschaft" within the SFB-386 ("Analysis of discrete structure," project B10 "Genetic epidemiological methods: Analysis of haplotype risk structure"), the German National Genome Research Net, the "Austrian Genome Research Programme" and the Munich Center of Health Sciences.

References

Akey, J., Jin, L. & Xiong, M. (2001) Haplotypes vs single marker linkage disequilibrium tests: What do we gain? *Eur J Hum Genet* **9**, 291–300.

Carroll, R. J., Ruppert, D. & Stefanski, L. A. (2006) *Measurement error in nonlinear models*. Boca Raton, FL: Chapman & Hall/CRC.

Clark, A. G. (2004) The role of haplotypes in candidate gene studies. *Genet Epidemiol* **27**, 321–333.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. (2001) High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229–232.

Epstein, M. P. & 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. & Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**, 921–927.

Govindarajulu, U. S., Spiegelman, D., Miller, K. L. & Kraft, P. (2006) Quantifying bias due to allele misclassification in case-control studies of haplotypes. *Genet Epidemiol* **30**, 590–601.

Heid, I. M., Wagner, S. A., Gohlke, H., Iglseder, B., Mueller, J. C., Cip, P., Ladurner, G., Reiter, R., Stadlmayr, A., Mackevics, V., Illig, T., Kronenberg, F. & Paulweber, B. (2006) Genetic architecture of the APM1 gene and its influence on adiponectin plasma levels and parameters of the metabolic syndrome in 1,727 healthy caucasians. *Diabetes* **55**, 375–384.

Heid, I. M., Lamina, C., Küchenhoff, H., Fischer, G., Klopp, N., Kolz, M., Grallert, H., Vollmert, C., Wagner, S., Huth, C., Müller, J., Müller, M., Hunt, S. C., Peters, A., Paulweber, B., Wichmann, H. E., Kronenberg, F. & Illig, T. (2008) Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *Am J Epidemiol* **168**, 878–889.

Johnson, G. C. L., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C. J., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, SCL., Clayton, D. G. & Todd, J. A. (2001) Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233–237.

Kraft, P., Cox, D. G., Paynter, R. A., Hunter, D. & De, V. I. (2005) Accounting for haplotype uncertainty in matched association studies: A comparison of simple and flexible techniques. *Genet Epidemiol* **28**, 261–272.

Kuchenhoff, H., Mwalili, S. M. & Lesaffre, E. (2006) A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics* **62**, 85–96.

Lake, S. L., Lyon, H., Tantisira, K., Silverman, E. K., Weiss, S. T., Laird, N. M. & Schaid, D. J. (2003) Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* **55**, 56–65.

Lamina, C., Bongardt, F., Kuechenhoff, H. & Heid, I. M. (2008) Haplotype reconstruction error as a classical misclassification problem. *PLoS ONE* **3**, e1853.

Levenstien, M. A., Ott, J. & Gordon, D. (2006) Are molecular haplotypes worth the time and expense? A cost-effective method for applying molecular haplotypes. *PLoS Genet* **2**, e127.

Mensah, F. K., Gilthorpe, M. S., Davies, C. F., Keen, L. J., Adamson, P. J., Roman, E., Morgan, G. J., Bidwell, J. L. & Law, G. R. (2007) Haplotype uncertainty in association studies. *Genet Epidemiol* **31**, 348–357.

Morris, R. W. & Kaplan, N. L. (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* **23**, 221–233.

Moskvina, V. & Schmidt K. M. (2006). Susceptibility of biallelic haplotype and genotype frequencies to genotyping error. *Biometrics* **62**, 1116–1123.

Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. (2005) Genotyping errors: Causes, consequences and solutions. *Nat Rev Genet* **6**, 847–859.

Ranade, K., Chang, M. S., Ting, C. T., Pei, D., Hsiao, C. F., Olivier, M., Pesich, R., Hebert, J., Chen, Y. D., Dzau, V. J., Curb, D., Olshen, R., Risch, N., Cox, D. R. & Botstein, D. (2001). High-throughput genotyping with single nucleotide polymorphisms. *Genome Res* **11**, 1262–1268.

Schaid, D. J. (2004) Evaluating associations of haplotypes with traits. *Genet Epidemiol* **27**, 348–364.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* **70**, 425–434.

Spinka, C., Carroll, R. J. & Chatterjee, N. (2005) Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genet Epidemiol* **29**, 108–127.

Stephens, M., Smith, N. J. & Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978–989.

Stram, D. O., Haiman, C. A., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E. & Pike, M.C. (2003) Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum Hered* **55**, 27–36.

Wong, M. Y., Day, N. E., Luan, J. A. & Wareham, N. J. (2004). Estimation of magnitude in gene-environment interactions in the presence of measurement error. *Stat Med* **23**, 987–998.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M.J. & 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. & 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.

Zhu, W. & Guo, J. (2006) A likelihood-based method for haplotype association studies of case-control data with genotyping uncertainty. *Sci China A Math* **49**, 130–144.

Supporting Information

Additional supporting information may be found in the online version of this article:

Figure S1 Mechanism of the MC-SIMEX approach: The naive estimator for $\lambda = 0$ and estimators calculated from simulated data with additional artificial misclassification ($\lambda > 0$) are plotted. The fitted curve (solid line) is extrapolated back to $\lambda = -1$ (dashed line), resulting in the MC-SIMEX estimator.

Table S1 Misclassification matrices for *APM1* haplotypes.

Table S2 Frequency, sensitivity, and specificity for *APM1* haplotypes.

Table S3 Performance of MC-SIMEX correction of haplotype misclassification in simulation studies (including results of MC-SIMEX estimators using linear and quadratic extrapolation functions).

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organised for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received: 14 September 2009
Accepted: 29 April 2010