From
the Institute of Medical Information Processing, Biometry, and Epidemiology
of the University of Munich, Germany
Chair of Epidemiology: Prof. Dr. med. Dr. rer. nat. H.-Erich Wichmann
and
the Institute of Epidemiology,
Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH)
Director: Prof. Dr. med. Dr. rer. nat. H.-Erich Wichmann

Misclassification in genetic variants and its impact on genetic association studies

Thesis Submitted for a Doctoral degree in Human Biology
at the Faculty of Medicine Ludwig-Maximilians-University,
Munich, Germany

by
Claudia Lamina
from
Augsburg, Germany
2009

With approval of the Medical Faculty
of the University of Munich

First reviewer:	Prof. Dr. med. Dr. rer. nat. H.-Erich Wichmann
Second reviewer:	Priv. Doz. Dr. Roland Kappler
	Priv. Doz. Dr. Rudolf A. Jörres
Co-supervision:	Dr. rer. biol. hum. Iris Heid
Dean:	Prof. Dr. med. Dr. h.c. M. Reiser, FACR, FRCR
Date of the oral examination:	28.04.2009

Acknowledgment

First of all, I would like to thank Prof. Dr. Dr. H.-Erich Wichmann, for the opportunity to work on this interesting and varied topic and for the constant support. I am grateful for the inspiring working environment at the Institute of Epidemiology and the manifold chances to present this work at national and international conferences.

I would like to thank Dr. Iris Heid, who closely supervised this work and enriched it with many fruitful ideas, competent advices and discussions. Many thanks for the support at all times.

Many thanks to all the people who contributed either with data or ideas, above all PD Dr. Thomas Illig, head of the working group 'Molecular Epidemiology' and co-interim head of the working group 'Genetic Epidemiology' at the Institute of Epidemiology, and Prof. Dr. Florian Kronenberg, Head of the Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University.

I also would like to thank Friedhelm Bongardt for his preliminary work and good cooperation.

Furthermore, I thank all my colleagues at the Institute of Epidemiology for their contributions on the excellent working conditions. Many thanks particularly to Martina Müller for her invaluable help on statistical and programming problems and many productive discussions.

Table of Contents

1. Introduction

The 20th century has seen a burst of knowledge and technologies in genetics and genetic epidemiology. Starting from the Mendelian laws, that where discovered in the beginning of the 20th century, the basis for modern molecular genetics was set in the 1950s, when Watson and Crick found the double helix structure of the DNA. What followed was gaining insights in the synthesis of proteins from genes and thus the key concept of genetics. With the first step of completing the DNA sequence description in 2001 [Lander et al., 2001;Venter et al., 2001], a map of bases on chromosome strands of the human genome has been presented, still leaving open the identification of genes determining or enhancing disease development. Thus, "this is just halftime for genetics", as Eric Lander, one of the fathers of the Human Genome Project, stated in 2001.

Diseases, that are caused by alterations in one single gene, are called monogenic diseases. Generally, those diseases are with some exceptions very rare and their inheritance mode follows Mendelian laws. Therefore, the genetic basis was first discovered for these monogenic disorders.

However, in the most cases where diseases are caused or altered by genes, the relation is more complex and based on many genes, which can also influence each other. The detection of the causes and the pathway of these complex human diseases is one of the next big goals in human genetic research. Genetic complex diseases do not follow a clear inheritance mode. They are characterized by being caused jointly by an unknown number of genetic variants, many environmental factors and their interactions and thus are also called multifactorial diseases. Examples for such diseases are diabetes mellitus, myocardial infarction, asthma or cancer. Due to their high prevalence in the population their relevance for the public health system is enormous. Unraveling the genetic mechanisms of such complex diseases is a difficult task, requiring methods from many different scientific fields, including genetics, biology, medicine, epidemiology and statistics.

It is common for most complex phenotypes, that they have been the objective of classical epidemiological research before genetics came into play. In the 1980s the field of genetic epidemiology was established as a conglomeration of classical epidemiology, molecular genetics, population genetics, statistics and bioinformatics. One of the first definitions of Genetic Epidemiology was given by Newton E. Morton [Morton, 1982], defining it as "a science which deals with the etiology, distribution and control of diseases in groups of relatives and with inherited causes of diseases in populations".

At the beginning, genetic epidemiological research was based primarily on family studies applying segregation and linkage analysis. With the emerging focus on complex diseases, more and more studies were then planned and conducted on unrelated subjects due to high prevalence in the population. Once it is clear, that there is a heritable component to a disease, classical epidemiological methods can be used together with methods that are essential to incorporate genetic factors. These specific components to genetic epidemiology involve different kinds of variants, like SNPs (Single Nucleotide Polymorphisms, section 1.1.1) or haplotypes (section 1.1.3), with the aim of establishing a genetic association of these with a specific disease.

In the following introductory sections, these kinds of genetic variants and their statistical modeling in genetic association studies are explained. Furthermore, a special focus is set on the problem of misclassification in general and in the genetic epidemiological setting in particular, which lays the ground for the methods described in the main part of this work on "Misclassification in genetic variants and its impact on genetic association studies".

1.1. Genetic variants

The genetic code in the human genome is specified by four DNA bases, the nucleotides Adenine (A), Cytosine (C), Guanine (G) and Thymin (T). More than 99% of the nucleotides in the DNA are the same from person to person and thus *monomorph*. DNA can vary in single bases, and therefore be polymorph at these special sites, or in the number of DNA sequences, which are repeated (e.g. microsatellites) (for details see [Thomas, 2004] or [Bickeböller and Fischer, 2007]). In the following, the most common variants used in genetic association analyses, single nucleotide polymorphisms (SNPs) and haplotypes, are described.

1.1.1. Single-Nucleotide Polymorphisms (SNPs)

Single Nucleotide Polymorphisms (*SNPs*) are DNA loci that vary from person to person in one base-pair, as shown in Figure 1. The possible nucleotides that are present in one population at a specific locus are called *alleles*, and the combinations of alleles from both chromosomes in one person are called *genotypes*. That is, genotypes are the realizations of one SNP for each single person. For person 1 (Figure 1), for example, the alleles A and T can be found on the first SNP with genotype A/T. This SNP exhibits the possible genotypes A/A, A/T and T/T in the population. For person 1, there are two different alleles (A and T) at the

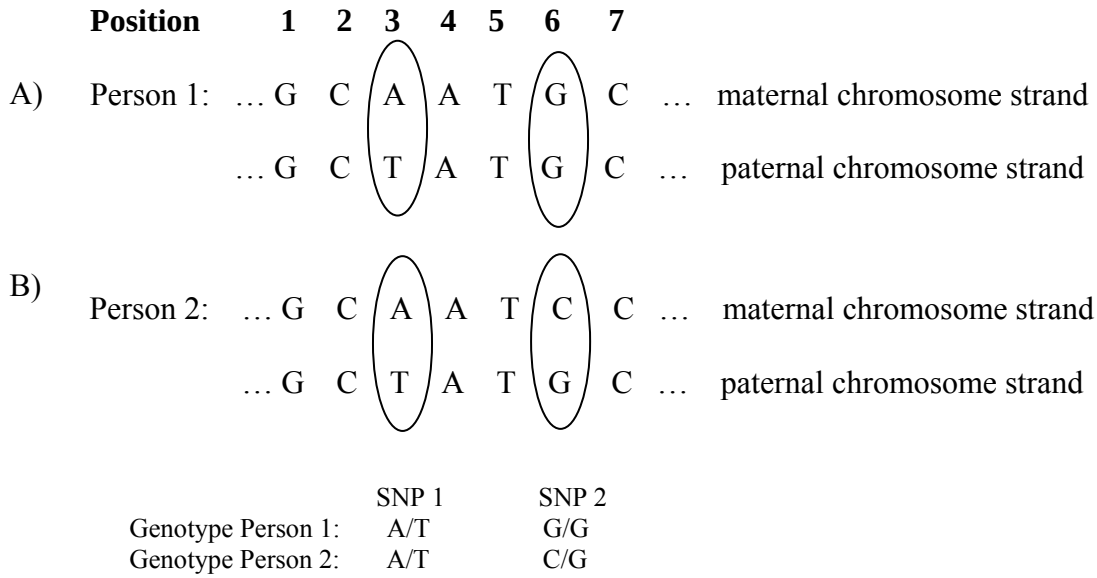


Figure 1: *One strand of each of the homologous chromosomes with SNPs on position 3 and 6 with genotypes A/T and G/G for person 1 (A) and A/T and C/G for person 2 (B)*

two chromosomes, which is called *heterozygous* at this locus. For two copies of the same alleles (in this case A/A or T/T), the person is called *homozygous* at this locus, as it is the case for the second SNP for person 1. Markers with exactly two possible alleles, as it is the case for almost all SNPs, are called *biallelic*.

SNPs are DNA variation in which each possible variant is present in at least 1 percent of people in a population by definition. Less common variations are called *mutations*.

SNPs occur with an average distance of about 300 base pairs. Therefore about 10,000,000 SNPs can be expected in the human genome.

SNPs are found all throughout the genome. They can have functional consequences by leading to changes in amino acid sequences in a gene or by alterating regulatory mechanisms of a gene in regulatory or intronic regions of a gene. Most SNPs, however, are found outside of the genes, in intergenic regions. Their functional consequences are not clear, yet, and have to be evaluated. However, SNPs, that have been found to be associated with a disease, don't have to be causal by themselves. They can also serve as markers for unmeasured, but correlated SNPs.

In association analysis, SNPs are the most popular variant due to their availability in high throughput technologies. Nowadays, microarray-based technologies are available which determine up to 1,000,000 SNPs efficiently and in an appropriate timeframe. Information of sequences and frequencies in different populations are collected in open databases (e.g.

dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) and HapMap (http://www.hapmap.org)), which makes them easily accessible.

For many statistical methods, assumptions have to be made on their distribution within each population. Assuming a large homogeneous randomly mating population, the genotype distributions of alleles A and a with respective allele frequencies $P(A) = p$ and $P(a) = q$, reach a balance after several generations, the so called *Hardy-Weinberg Equilibrium* (*HWE*): $P(A/A) = p^2$, $P(A/a) = 2pq$, $P(a/a) = q^2$. Deviation of HWE can be tested by comparing observed genotype frequencies with the expected under the above stated assumption via a χ^2-test. Such a deviation might be due to measurement error or the existence of unknown subpopulations.

1.1.2. Linkage Disequilibrium

Correlation between SNPs is expressed by the concept of *Linkage Disequilibrium*. If two loci are far apart, chances are high that there has been one or several crossover of homologous chromosomes (*recombination*) and thus inheritance of the allele on one locus is rather independent from the allele on the second locus. If there is high *linkage* between two loci, the alleles on both loci are mostly inherited jointly, and thus are called to be in *linkage disequilibrium* (*LD*). In this case, the observed frequency of the joint presence of two alleles differs from the expected frequency assuming independence between these SNPs. If loci are in high LD, one SNP that is genotyped in the laboratory may serve as a marker for another SNP that is not genotyped, but may be the causal variant. Common LD measurements are e.g. *Lewontins D'* [Lewontin, 1964] or the *correlation coefficient* r^2 [Devlin and Risch, 1995] If D' between two loci equals 1, there is no indication for recombination between them. However, that doesn't imply that these two loci carry the same information. This situation is illustrated in Figure 2a): The G-allele in locus 2 can be predicted by the C-allele in locus 1, while the T-allele in locus 1 can be predicted by the A-allele in locus 2, but not the other way round. Thus, the information is not redundant. Figure 2b) shows an example for r^2=1. The two loci provide the same information. This is exactly the case, if D' is 1 and the allele frequencies of both SNPs equal each other [Cordell and Clayton, 2005].

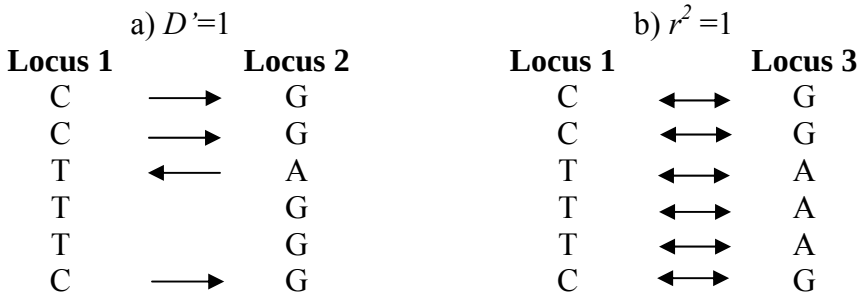


Figure 2: *Example for LD between SNPs: a) D'=1 between Locus 1 and Locus 2 and b) r^2 =1 between Locus 1 and Locus 3, with arrows showing the direction in which an allele can be predicted from the other one.*

Recombination does not take place equally distributed within the genome, but there are blocks, that are preserved over many generations and spots in-between where a lot of recombination occurs (recombination hot spots) [Daly et al., 2001;Gabriel et al., 2002]. Within these blocks, SNPs are highly correlated. That is, some SNPs within those LD-blocks may be redundant and can be omitted in the modeling process without loss of information. Thus, the dimension of problems involving these SNPs can be reduced, e.g. by estimating haplotypes.

1.1.3. Haplotypes

Haplotypes have been the subject of considerable attention as they complement the information from the SNP genotypes. The nucleotides of the two DNA strands of one person can be stringed together as two long rows of code as in Figure 1. Genotypes summarize the vertical information on a single locus across the two strands while haplotypes summarize the horizontal multi-locus information per strand. That is, haplotypes are sequences of alleles inherited from one parent. Since monomorph loci do not give additional information, haplotypes are only based on polymorphic loci (SNPs).

Figure 3 shows the haplotypes based on two SNPs for two different persons. The haplotypes based on SNP 1 and SNP 2 as shown in Figure 1, are thus AG and TG for person 1 and AC and TG for person 2.

Genotypes can be deduced uniquely from haplotypes, if they are known. Haplotypes hold additional *phase* information compared to genotypes, namely the information, which alleles are inherited jointly on one chromosome strand. Therefore, haplotypes can be deduced uniquely from genotypes only for those persons, where all SNPs that are included in the haplotype are homozygous, except at most one SNP.

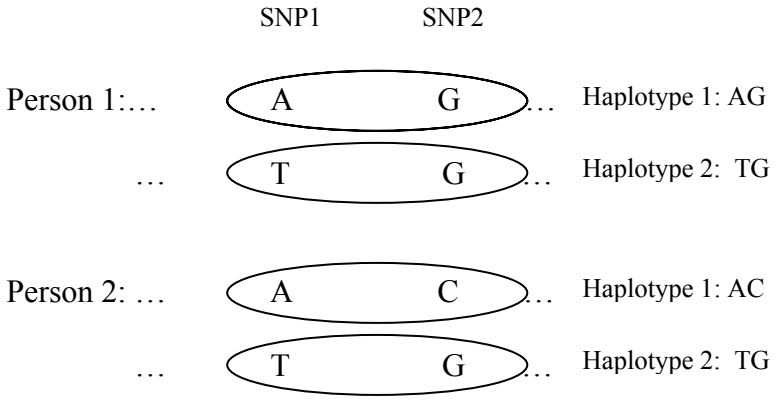


Figure 3: *Haplotypes based on SNP 1 and SNP 2 for persons 1 (observed haplotypes: AG and TG) and person 2 (observed haplotypes: AC and TG)*

Due to their phase information, there are several advantages of haplotypes:

1. In regions of high linkage disequilibrium, haplotype diversity is limited due to high correlation between the SNPs resulting in only few existing haplotypes in the population [Daly et al., 2001;Gabriel et al., 2002]. Thus, haplotypes, which hold information on more than one locus, may capture the LD information in a gene better than methods based on single loci [Akey et al., 2001a]. This results in a gain of power compared to an analysis which incorporates all existing single SNPs [Akey et al., 2001a;Morris and Kaplan, 2002].
2. Generally, the diversity of haplotypes is captured by a few common haplotypes and some rarer ones. The majority of the diversity within a region can be picked up by typing those SNPs which explain a high percentage of variance (e.g. more then 80 or 90 percent) in haplotypes [Stram et al., 2003a], called *tagging* or *tag SNPs*. Figure 4 shows, how redundant information can be removed, reducing the overall costs of genetic association studies: The tagging SNPs 1, 3 and 4 sufficiently represent the haplotype diversity in this region of seven SNPs.
3. The information on the causal variant that may not be genotyped directly can be captured by haplotypes due to the correlation structure within one LD-block [Clark, 2004;Schaid, 2004]. In Figure 4, for example, the information of SNPs 2, 5, 6 and 7 is captured from the haplotypes, even though they might not be genotyped.

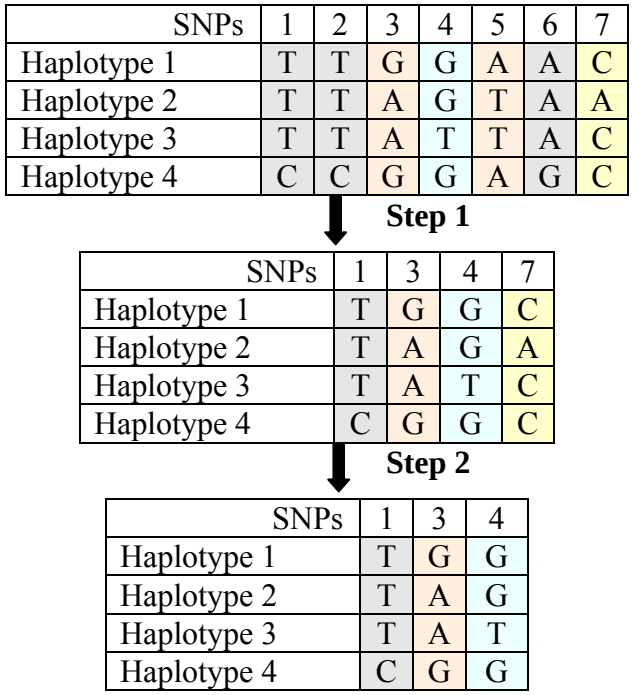


Figure 4: *In a region of seven SNPs four haplotypes are present in the study. SNPs 1, 2 and 6 provide exactly the same information for the haplotypes, as well as SNPs 3 and 5. Thus, three SNPs can be removed in step 1. In step 2, SNP 7 can be removed, since it doesn't give additional information to the combination of alleles from SNP 1,3 and 4, leaving SNPs 1,3 and 4 as tagging SNPs for all haplotypes in this region.*

4. Finally, the haplotype may represent the biologically functional genetic unit better than the genotypes [Clark, 2004]. Haplotypes can thus provide additional information with respect to association analysis and localization of complex disease genes [Martin et al., 2000], especially in the presence of more than one susceptibility allele [Morris and Kaplan, 2002].

Therefore, haplotype association analyses can add substantially to association analyses based on SNP genotypes.

1.2. Genetic association studies

1.2.1. The aim of genetic association studies

The hypothesis that specific genetic variants, like SNPs or haplotypes, are associated with a phenotype of interest is tested in genetic association studies. Regardless of the definition of this outcome as a disease, a disease-related marker or even a non-disease related trait, the

outcome variable to be explained by genetic and environmental variables is called *phenotype* in the following. As shown in Figure 5, genetic complex diseases are caused by a complex mechanism of many genetic and non-genetic factors and their interactions. They can modify either other risk factors, also called intermediate phenotypes, or the disease itself. Since association analyses are based on frequencies of genetic variants in the study population, the detected association does not have to be causal but can be due to correlation of the specific variant with a truly near-by located causal factor that is not measured.

Genetic association studies are essential tools in finding the pathway between a genetic variant and a phenotype. To establish reproducible results, large datasets, independent replication of results and appropriate usage of genetic epidemiological methods are necessary.

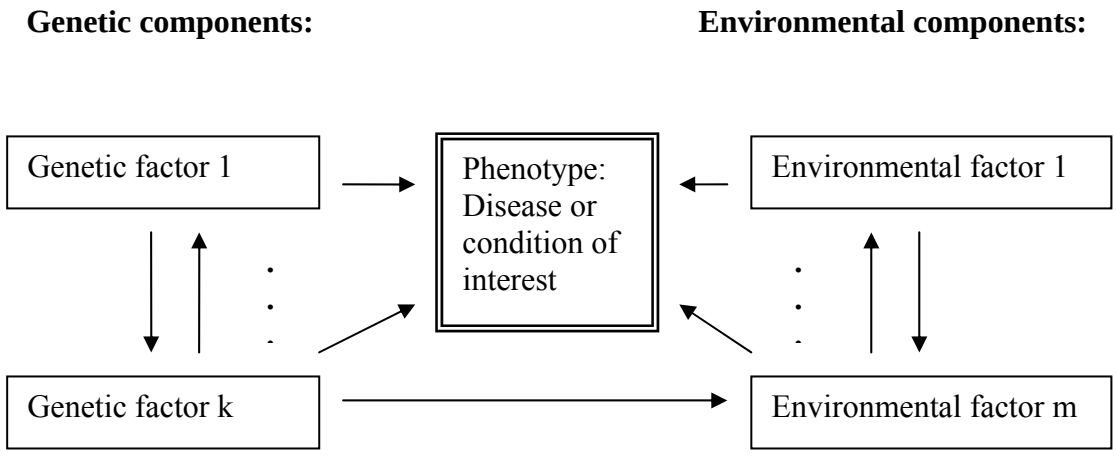


Figure 5: *Schematic model of etiology mechanisms in complex genetic diseases*

1.2.2. Approaches to identify genes

Genetic association studies have been conducted successfully to analyze regions in the genome that were a priori assumed to be associated with the disease of interest. Those candidate gene regions can be identified by hypothesis-driven as well as hypothesis-free approaches:

1. The candidate gene may have an already known biological function that can plausibly be related to the disease. All genes and variants, that play a role in the assumed pathophysiology of the disease, can be seen as candidate genes. For example, for coronary artery disease, enzymes involved in cholesterol metabolism might be promising candidate genes.

2. Results of *linkage studies*, a family-based study design [Thomas, 2004], may give hints on several gene regions.
3. Variants associated with the disease can be identified by *Genome Wide Association studies* (*GWA*) on so far up to 1,000,000 SNPs distributed over the whole genome [McCarthy et al., 2008]. Genome wide association studies do not differ from candidate gene approaches with respect to study design or analysis methods. However, it is a hypothesis-free shotgun method leading to many false positive results, while a candidate gene association study looks at preselected gene regions and variants. Therefore, results of GWAs have to be accompanied by replication studies in other populations followed by functional studies and/or, animal models.

After gene regions have been selected as candidate genes, it is then the purpose of candidate gene association studies to validate the association of the candidate gene with the disease, to narrow down this region and to pinpoint the responsible variant by fine-mapping techniques.

1.2.3. The problem of non-replication in genetic association studies

Genetic association studies have proven to be successful tools by providing reliable and replicable association results of genetic variants with e.g. diabetes mellitus [Altshuler et al., 2000;Helgason et al., 2007;Zeggini et al., 2007], Crohns disease [Economou et al., 2004], breast cancer [Stacey et al., 2008] or coronary disease [Helgadottir et al., 2007;McPherson et al., 2007;Schunkert et al., 2008;Ye et al., 2008]. However, replication of consistent results is often difficult [Chanock et al., 2007;Hirschhorn et al., 2002;Redden and Allison, 2003].

There are several reasons for non-replication in genetic association studies. Most effects of single genetic variants are expected to be very low. Many studies are too small and therefore underpowerd to find these small effects. Meta-analyses may overcome this situation. However, there might be an additional problem with heterogeneity between studies [Ioannidis et al., 2007] and/or insufficient phenotyping or differing phenotype definitions between studies.

In the era of genomewide association studies, there is an inflating number of genetic association tests conducted, lifting the multiple testing problem on a very high scale. The more statistical tests are conducted, the higher are chances for significant findings by chance alone. Therefore, many replication studies with „negative“ findings are expected and even necessary to seperate the wheat from the chaff [Ioannidis, 2007]. Thus, accounting for

multiple testing is essential, especially for genome-wide association studies, but also for candidate gene studies, where more than one genetic variants are involved.
Non-replication might also be due to population stratification, i.e. the occurrence of undetected subpopulations [Lamina et al., 2005;Steffens et al., 2006], measurement errors in phenotypes or misclassification in explaining variables as in genotypes or haplotypes.

1.2.4. Generalized linear models

Genetic association studies are most commonly designed as case-control or cross-sectional studies, which are standard approaches in epidemiology [Kreienbrock and Schach, 2005]. The phenotype of interest may thus be a dichotomous or a quantitative variable. Standard methods to analyze case-control or cross-sectional studies can also be applied for genetic association studies.
The most popular approach to analyze genetic association is to array the data of the case-control status versus the genetic variant in an ordinary contingency table. If the frequency of a specific genetic variant in the cases is higher or lower than would have been expected by chance alone, an association is detected. The statistical significance of the differences between allele frequencies (2x2 table) or genotype frequencies (2x3 table) can then be assessed by a χ^2-Test. If an ordering of the genotypes can be assumed (see section 1.2.5), an Armitages trend test can be conducted. The analog in the quantitative case would be a t-Test on differences of the mean between genotype, allele or haplotype values.

In the framework of *Generalized Linear Models* (*GLM*) [McCullagh and Nelder, 2008], additional covariates and complex genetic models or interaction can be taken into account.
For a quantitative outcome variable *Y*, a multiple linear regression model describes the association of *k* observed covariates $X_1, \ldots X_k$ in the linear term

$$Y = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \ldots + \hat{b}_k X_k + \varepsilon ,$$

with ε being a normally distributed error term ($\varepsilon \sim N(0,\sigma^2)$) and $(\hat{a},\hat{b}_1,\hat{b}_2,\ldots\hat{b}_k,\sigma^2)$ being unknown parameters, that have to be estimated from the data. The covariates explain some proportion of the outcome variable and are thus also called explanatory or independent variables.
For a dichotomous phenotype (e.g. *Y=1*: Person is affected, *Y=0*: Person is not affected), the logistic regression model can be formulated as:

$$\text{logit}(Y) = \ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + ... + \hat{b}_k X_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

P(Y=1) and *P(Y=0)* define the probabilities, that a person is affected or unaffected. For both, linear and logistic regression, the β-parameters can be interpreted as the slopes of the linear line from *X* to *Y* or logit(*Y*), respectively. For zero slopes, there is no association between this covariate and the outcome. Therefore, tests of significance (e.g. Wald Test) test the following hypothesis: Nullhypotesis $H_0 : \beta = 0$ versus the alternative $H_1 : \beta \neq 0$

In the case of a simple linear or logistic regression model with just one covariate, this test equals the t-Test or χ^2-Test.

Commonly, variables like gender or age are taken as covariates in epidemiological studies to reduce the phenotypic variance. Other covariates may be chosen as they may act as confounders in the association analysis. In genetic association studies, the genetic variants of interest are also included into the model as covariates. They have to be coded based on the assumed genetic model between the genetic variant and the outcome variable.

1.2.5. Modeling genetic effects

For one SNP with two alleles, there are three possible genotype values: *AA, aA* and *aa,* with *a* being the minor, that is the less frequent allele, and *A* being the major allele. The distribution of a quantitative phenotypic trait *Y* or the probability of exhibiting one particular trait *Y=1* given a specific genotype value is expressed by $P(Y \mid AA)$, $P(Y \mid aA)$ and $P(Y \mid aa)$ and can be modeled in four different ways:

1. *Dominant*: $P(Y \mid aA) = P(Y \mid aa)$

   Subjects with one copy of the minor allele have the same „risk“ exhibiting the phenotypic trait than subjects with two copies.

   Covariate *X* is coded as follows: *X=1*, if genotype = *aa* or *aA*

   X=0, if genotype = *AA*

2. *Recessive*: $P(Y \mid aA) = P(Y \mid AA)$

   Only subjects with two copies of the minor allele exhibit the phenotypic trait.

   Covariate *X* is coded as follows: *X=1*, if genotype = *aa*

   X=0, if genotype = *aA* or *AA*

3. *Additive*: $P(Y \mid aA)$ lies exactly between $P(Y \mid aa)$ and $P(Y \mid AA)$. The risk is increasing linearly for each copy of the minor allele.

Covariate X is coded as follows: $X=2$, if genotype = aa

$X=1$, if genotype = aA

$X=0$, if genotype = AA

4. *Codominant / Model-free*: $P(Y \mid aa) \neq P(Y \mid aA) \neq P(Y \mid AA)$

   There is no restriction on the size or direction of the genetic effect on the phenotype.

   Covariate X is coded as follows: $X_1=1$, if genotype = aa

   $X_1=0$, if genotype = aA or AA

   $X_2=1$, if genotype = aA

   $X_2=0$, if genotype = aa or AA

Therefore, given the codominant model, two parameters are estimated for one SNP. The distribution of a normally distributed quantitative trait given hypothetic genetic variants for dominant, recessive and additive effect models is given in Figure 6.

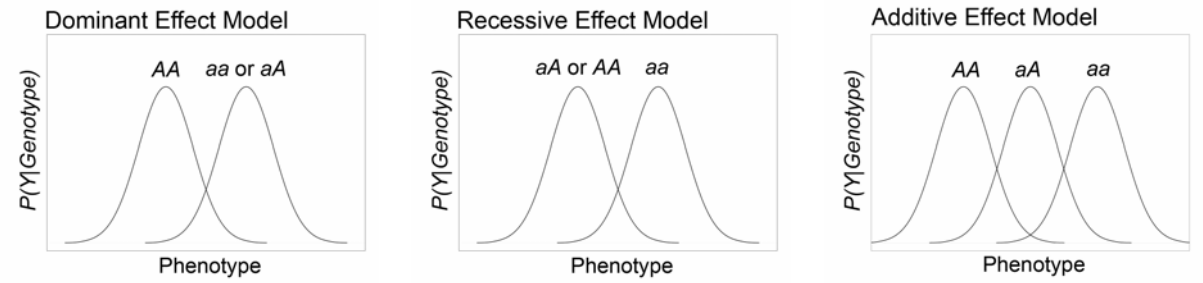


Figure 6: *The distribution of a normally distributed quantitative trait given genetic variants for dominant, recessive and additive effects*

This distribution or the probability of a trait given haplotypes can be expressed in the same way by modeling 0, 1 or 2 copies of a particular haplotype in a dominant, recessive, additive or codominant way as for the number of minor alleles for a SNP-variable as defined above.

1.2.6. Example: The *APM1* gene and its association with adiponectin plasma levels

SNP genotypes and haplotype data in the SAPHIR Study (Salzburg Atherosclerosis Prevention Program in Subjects at High Individual Risk, [Heid et al., 2006]) are shown as a representative example of successful association studies. This is an observational study conducted in 1999–2002 involving 1,770 healthy unrelated subjects of Caucasian origin between 40 and 70 years of age. Originally, 53 SNPs of the *APM1* gene, encoding for

adiponectin, were selected and genotyped in a subsample of 81 persons. To reduce genotyping costs, haplotype tagging SNPs were selected out of these (using the program tagSNP [Stram et al., 2003a]). These tagging SNPs were then genotyped in the whole smple of 1770 persons. This strategy resulted in 15 haplotype tagging SNPs with minor allele frequencies of >1%.

The primary aim of this study was to detect association with plasma adiponectin levels. A linear regression model was applied with the transformation log(adiponectin+1) as the quantitative outcome variable. Association was tested for each SNP separately, applying an additive as well as codominant model.

Haplotypes were reconstructed based on the 15 selected haplotype tagging SNPs via the EM algorithm (SAS proc haplotype). For haplotype association analysis, all observed haplotypes with frequency of >1% were included in one regression model, except the most common haplotype, which served as a reference group.

All SNP- and haplotype association analyses were adjusted for age, sex and BMI.

Strong associations with adiponectin plasma concentrations were found for 11 of the 15 SNPs and for 9 of the 18 observed haplotypes, indicating a clear modulation of adiponectin concentrations by *APM1* variants.

However, most observed haplotypes in the *APM1* gene are rather rare, primarily the haplotypes with the strongest associations with a frequency of 5.3% (haplotype H2) and 2.3% (haplotype H12). It might be discussed, that rarer haplotypes are more likely to be erroneous than more common haplotypes. Therefore, knowledge of the size and structure of misclassification in genotypes and haplotypes and its impact on association estimates would be helpful for a thorough interpretation of association results in this study.

1.3. Misclassification as a statistical challenge

1.3.1. Estimating misclassification of independent variables

In each situation, where a variable is measured or sampled, errors may arise. For quantitative values, this error is called *measurement error* [Carroll et al., 2006]. *Misclassification* occurs, if a measured or observed category of a discrete variable differs from the true category [Gustafson, 2003]. These situations can arise through erroneous measurements, like categorizing a patient into being hypertonic, while being normotonic in reality after a blood pressure measurement or through a wrong answer in a questionnaire.

This magnitude of misclassification is expressed through classification probabilities, answering the question, "How likely is the correct classification given the true classification?". For the dichotomous explanatory variable X (i.e. the truth), which is substituted through the imperfect surrogate X* (i.e. the observed), the misclassification is described by the concept of sensitivity and specificity. The *sensitivity* defines the probability that a true positive is correctly classified ($P(X^* = 1 \mid X = 1)$) while the *specificity* expresses the probability that a true negative is correctly classified ($P(X^* = 0 \mid X = 0)$).

For discrete explanatory variables with more than two categories, sensitivity and specificity are not sufficient to explain the misclassification scheme. In this case, misclassification probabilities can be expressed via a misclassification matrix with $\pi_{ij} = P(X^* = i \mid X = j)_{i,j}$ being the probability that the category *i* is observed, while *j* is the true category.

Since the true values are not known, misclassification probabilities cannot be calculated directly, but can be estimated by several different approaches, either internally or externally. For internal validation methods, the extent of error is evaluated within the study data at hand, while external validation is based on data of other but comparable studies. For both methods, the extent of misclassification can be estimated by:

1. Comparison to a gold standard: In the best of cases, there is a *gold standard* available for a subset of the data (internal method) or for another, possibly smaller study (external method). A gold standard is a sampling or measurement method which can be assumed to be without errors, but that is too costly or complicated to be applied on the full sample. By comparing the measurements from the gold standard to the measurement technique used in the study, misclassification probabilities can be estimated.
2. Calculation from replicate measurements: Another method to estimate these probabilities are based on replicated measurements. Since all measurements are possibly error-prone, misclassification cannot be calculated from multiple measurements directly, but the number of discordances within the replicated measurements can be counted. Maximum likelihood methods can then be used to estimate misclassification probabilities.
3. Approximation by simulation: If methods 1 and 2 cannot be applied, error measures and misclassification probabilities can be approximated via simulations: Data are simulated, which serve as the true values, and measured or observed in the same way as the variable of interest. In the simulated data, the true and the observed values are known, so that misclassification probabilities can be calculated directly. These

probabilities can then be taken as approximations for the misclassification in the study data.

Besides the quantification of the size of misclassification, its structure has to be characterized. For most methods, *nondifferential* misclassification is assumed. That is, the probability of misclassification does not depend on the outcome variable. In the case of case-control studies, the misclassification scheme does not differ between cases and controls. Other misclassification patterns that depend on the values of the outcome variable are called *differential*.

1.3.2. The effect of misclassified independent variables on association estimates

Misclassification in independent variables of association analyses models are known to yield biased estimates and decrease power to detect an association [Carroll et al., 2006;Thomas et al., 1993]. Generally, the bias depends on the magnitude of the error: The smaller sensitivity and specificity, the higher the bias of regression coefficients.

Let's define β as the true but unknown parameter in the linear or logistic regression and β^* the estimated naive coefficient to the observed variable X*.

For dichotomous nondifferentially misclassified variables, misclassification always leads to underestimation of regression coefficients, i.e. they are biased toward the null ($|\beta^*| < |\beta|$) [Bross, 1954]. This can be interpreted as a flattened regression line due to misclassification in a simple linear regression. This attenuation of coefficients worsens with the severity of misclassification. For polychotomous variables, however, even in the case of nondifferential error, the impact of misclassification is not always predictable [Dosemeci et al., 1990]. Nevertheless, situations where plausible measurement error leads to bias of effects away from the null are rather unusual [Gustafson, 2003].

In the case of differential misclassification, the direction of bias is not predictable.

1.3.3. Methods to account for misclassification

1.3.3.1. Overview of methods correcting for misclassification

The correction of measurement error models have been treated extensively in the literature [Carroll et al., 2006]. General methods accounting for misclassification, however, are rather sparse.

In the simplest case of a binary explanatory and binary outcome variable, the data can be expressed via a misclassified contingency table. If the misclassification probabilities, in this case the sensitivity and specificity are known, direct correction of corresponding Odds Ratios for misclassification bias by the matrix method can be applied [Morrissey and Spiegelman, 1999]. In principle, if the misclassification probabilities are known, misclassified contingency tables can also be handled by the maximum likelihood or quasi-likelihood methods (Carrol 2005) or within a Bayesian approach for binary explanatory variables [Gustafson, 2003] or for ordinal regression [Mwalili et al., 2005].

In all of these methods addressed so far, nondifferential misclassification and prior information about the quantity and pattern of misclassification is needed.

Furthermore, these approaches cannot be generalized to more complex cases that can be modeled within Generalized Linear Models, like modeling a quantitative outcome variable or accounting for covariates.

1.3.3.2. The MC-SIMEX method

In the presence of measurement error of quantitative variables, the *Simulation Extrapolation method* (*SIMEX*, [Cook and Stefanski, 1994]) has become a powerful tool for correcting effect estimates in the framework of Generalized Linear Models. The idea is that additional measurement error is simulated and added to the observed values of the independent variable (SIMulation step). For these data, effect estimates are calculated and then this trend in estimates over the increasing measurement error levels is EXtrapolated back to the case of no measurement error to provide an unbiased estimate.

This approach was modified to suit the general misclassification problem for categorical data, called the *Misclassification-SIMEX* (*MC-SIMEX*) [Kuchenhoff et al., 2006]. To correct for misclassification using this method the misclassification matrix Π has to be known. Briefly, this approach, which is pictured in Figure 7, consists of two steps, the SIMulation step and the EXtrapolation step. In the simulation step, data are simulated with increasing

misclassification $\Pi^{1-\lambda}$ with $\lambda > 0$. Association estimates $\hat{\beta}_\lambda *$ are computed for each λ starting from the observed data set ($\lambda = 0$) and the observed association estimate $\hat{\beta} * = \hat{\beta}_0$. A function (linear, quadratic, or log-linear) is fitted to the β estimates and extrapolated back to the case of no misclassification $\Pi^{1-\lambda} = I$ for $\lambda = -1$ in the extrapolation step. The estimate $\hat{\beta}_{-1}$ yielded from this extrapolation is the SIMEX-corrected estimate, which was shown to be consistent. It can be applied for all Generalized Linear Models for any given misclassification matrix Π, for which $\Pi^{1-\lambda}$ exists for all $\lambda \geq -1$. This should be the case for most useful misclassification matrices. For 2x2 matrices, existence is ensured, if $\det(\Pi) > 0$. Sparse misclassification matrices, however, especially matrices containing the null, might not fulfill the criterion of being valid misclassification matrices for the MC-SIMEX method. In this case, a matrix is chosen, which is as similar as possible to the original matrix.

To apply the MC-SIMEX, nondifferential misclassification has to be assumed and prior information about the quantity and pattern of misclassification is required.

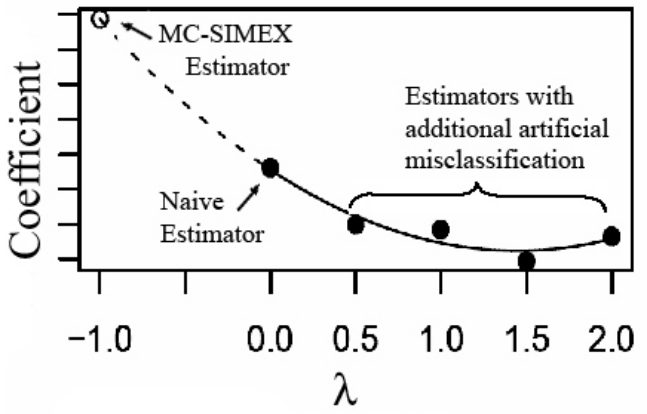


Figure 7: *Mechanism of the MC-SIMEX approach: The naive estimator for λ=0 and estimators with additional artificial association with λ>0 are plotted. The fitted line through these points is extrapolated back to λ=-1, resulting in the MC-SIMEX estimator.*

1.4. Misclassification in genetic variants

In genetic epidemiology, there is a high awareness of error sources and the necessity to account for it. However, only specific aspects of the misclassification problem on genotypes and haplotypes have been treated, so far. In the following, it is explained, why genotypes and haplotypes might be affected by misclassification. Previous works on these aspects are depicted with focus on arising problems and work that still lies ahead.

1.4.1. Misclassification due to genotype error

1.4.1.1. Genotype error sources

The process of data collection in genetic epidemiological studies usually involves the study center for phenotyping and blood withdrawal, genotyping lab for DNA processing and genotyping, as well as data management. In the study center, subjects are interviewed and the medical exam is taken including blood withdrawal. Blood is processed and put on mother plates. In the genotyping lab, DNA is processed, pipetted into plates, and SNPs are genotyped. The SNP data is entered into the data management system and merged with phenotypic information.

In each of these data collection steps, errors may occur by interactions between DNA molecules, sample quality issues, biochemical artifacts, equipment reliability or by the human factor [Pompanon et al., 2005]. For example, blood samples may be mismatched in the study center, in the genotyping lab, or by data management. Man-made or natural DNA deterioration may also lead to low quality DNA and thus faint error-prone genotyping signals.

1.4.1.2. Previous work on genotype error

Deriving the pattern and size of genotype error for unrelated subjects requires assumptions or validation and replication steps. Generally, genotype misclassification probabilities can be estimated either by obtaining measures of a gold standard in a subgroup as well as by obtaining repeated measurements (see section 1.3.1). Validation data implying the availability of a gold standard genotype would be ideal [Carroll et al., 2006] and was already proposed [Gordon et al., 2004], but may not be advisable as the perfect gold standard for measuring genotypes is not available yet. This approach might lead to over-correction when such a non-perfect standard is used [Wacholder et al., 1993].

Using replication from multiple genotype assessments (>2) was illustrated in small-scale experimental [Pompanon et al., 2005] or simulation data [Lai et al., 2007].

Instead of a gold standard method or multiple genotyping, double genotyping of 5-10% of a sample is often available from routine quality control. The discordance that can be calculated from these duplicate measurements as the number of discordant genotype pairs relative to the total number of pairs is often misnamed as the “genotyping error”. This is not correct as the discordance relates repeatedly observed values with each other, but not the observed with the

truth. The true genotyping error can be estimated from the discordance, but previous attempts on estimating the genotyping error involved only a limited number of discordants such as two [Tintle et al., 2007] or 30 [Wong et al., 2004] and implied restrictions on the error model.
Furthermore, realistic genotype error estimates from epidemiological samples as encountered in routine association analysis, are lacking. One reason might be methodological issues regarding identifiability of probability estimates. Another aspect might be that many researchers expect the error size to be too small to be of interest.
In contrast to that expectation of a small error, methodological investigations on the impact of genotyping errors often assume large errors (1-10%). Therefore, most simulations are not based on realistic assumptions.
There is numerous work on the effect of genotype error on linkage studies [Lincoln and Lander, 1992;Sobel et al., 2002], linkage disequilibrium [Akey et al., 2001b], the selection of tagging SNPs [Liu et al., 2006], Multiple Dimension Reduction Methods [Ritchie et al., 2003], genotype and haplotype distribution [Govindarajulu et al., 2006;Moskvina and Schmidt, 2006;Quade et al., 2005;Zhu et al., 2007], and on family-based association [Gordon and Ott, 2001;Mitchell et al., 2003;Morris and Kaplan, 2004;Seaman and Holmans, 2005].
The current work on the impact of genotype error on association studies on unrelated subjects, however, is mostly restricted to case-control studies using a χ^2-test or the Armitages trend test not allowing for covariate adjustment [Gordon et al., 2002;Gordon et al., 2004;Gordon et al., 2007;Gordon and Ott, 2001;Kang et al., 2004b;Rice and Holmans, 2003;Tintle et al., 2007]. There is few work for logistic [Lai et al., 2007] or linear [Wong et al., 2004] regression models incorporating only restricted error models.

1.4.2. Misclassification due to haplotype reconstruction

1.4.2.1. Haplotype error sources

The direct determination of haplotypes in the laboratory is time-consuming and still too expensive for a large number of individuals in epidemiological studies. High throughput technologies produce genotype data, from which haplotypes have to be inferred statistically. Haplotype reconstruction is unambiguous only for those persons which are homozygous at all SNPs that are involved in the haplotype or with at most one heterozygous SNP. For example, for person 1 (Figure 3) with genotypes A/T and G/G, there is only one possible haplotype combination. For person 2 however, with genotypes A/T and C/G, it is not clear only from the genotype data, if the A allele is on the same strand as the C or G allele. Therefore, there are

two possible combinations. AC and TG, as it is the truth in this example or AG and CT, which would be wrong. For these ambiguous cases, statistical reconstruction methods are needed to decide for one of the possible combinations.

The most frequently used methods for haplotype reconstruction are based either on the maximum likelihood-based expectation-maximization (EM) algorithm [Excoffier and Slatkin, 1995] or on a Bayesian framework incorporating the coalescent model [Stephens et al., 2001;Stephens and Donnelly, 2003] (see section 3.1.2). Whichever method is used, they all involve a certain amount of error in the reconstructed haplotypes.

Reconstructed haplotypes are based on measured genotypes. Thus, the possible error introduced through the genotypes (see section 1.4.1) is carried forward into the respective haplotypes estimated from these error-afflicted SNP genotypes. They are typically based on more than two and often more than 10 or even 20 SNPs. Therefore, the genotyping error accumulates over the number of SNPs involved. That is, the overall misclassification on haplotypes is always a combination of genotype error and the uncertainty due to statistical haplotype reconstruction, which is depicted in Figure 8.

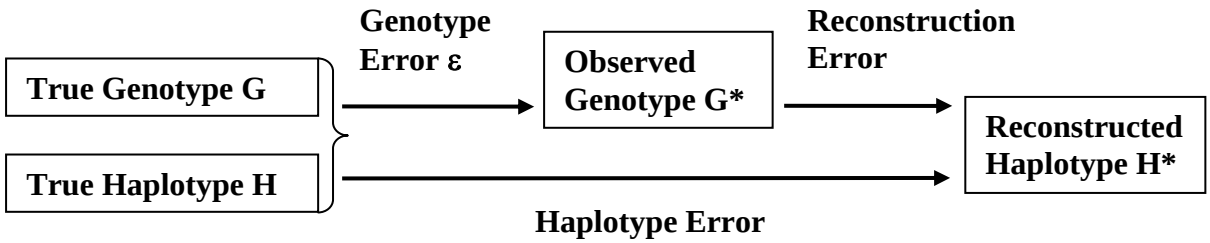


Figure 8: *Schematic overview of the haplotype error sources*

1.4.2.2. Previous work on Haplotype Error

Due to their statistical reconstruction, there is a certain amount of uncertainty involved in inferred haplotypes. Fallin and Schork [Fallin and Schork, 2000] investigated the haplotype error using the mean squared error (MSE), which was found to increase with increasing minor allele frequency (MAF), decreasing LD and increasing number of loci. However, the MSE summarizes the error in the estimated haplotype frequencies rather than the error in the individuals' haplotypes, which is of interest in haplotype association analyses. One commonly used error measure for the error in the individuals' haplotypes is the error rate of which some authors have described selected aspects [Adkins, 2004;Niu et al., 2002;Stephens et al., 2001;Xu et al., 2004] . However, a systematic investigation of the error rate is still lacking.

Furthermore, researchers are often interested in subjects carrying a specific haplotype and thus rather in the error in assigning this haplotype than in an error measure averaging across all haplotypes.

Besides work on haplotype error measures and the estimation of it, there is also some literature on the effect of haplotype errors:

The bias induced by pure reconstruction error on estimates from haplotype association studies has been shown to be non-negligible in simulation studies [Kraft et al., 2005]. There are retrospective as well as prospective likelihood methods, [Epstein and Satten, 2003;Lake et al., 2003;Schaid et al., 2002;Spinka et al., 2005], which account for this haplotype uncertainty by using estimates of the distribution of haplotypes given the observed genotypes to estimate haplotype risk parameters. These methods, though, are in most cases restricted to case-control studies or there are other limitations like not allowing for environmental covariates or restriction to additive effect models. In all cases, there is no possibility to account for genotyping error, yet. The error induced by the genotyping may accumulate across multiple loci, thus affecting haplotype association analysis. The impact of genotyping error on haplotype frequency was reported to be very small [Kang et al., 2004a;Zou and Zhao, 2003].

However, the uncertainties in the individually assigned haplotypes are of more immediate interest in many cases. For example, persons with certain haplotypes may be depicted for further in-depth studies to investigate metabolic changes by exposure to certain conditions. Those studies are expensive and thus, unambiguous or very certain haplotypes are preferred.

A more common problem is the impact of haplotype uncertainty on association analysis. The impact of genotyping error on haplotype association analysis in simulated case-control studies was reported to be non-negligible in situations of high haplotype complexity, high relative risks and high allele error rates [Govindarajulu et al., 2006]. However, unambiguous haplotype reconstruction has been expected. Therefore, the error induced by the uncertainty in the statistical reconstruction has not been accounted for.

A clear picture of the haplotype error resulting from a combination of reconstruction and genotype error and its impact on association particularly in real data situations is still lacking.

1.5. Objective of this investigation

The initial problem, which motivated this work, was the problem of non-replication in genetic association studies. The focus was set on misclassification in genetic variants, which might be one of the possible reasons.

Therefore, it is the objective of this work to quantify and characterize genotype and haplotype misclassification as possible sources for bias in genetic association studies.

To evaluate the dimension of this problem with respect to error sizes and impact on estimated effect parameters, three different aspects have to be considered:

1. The amount and structure of genotyping error,
2. The amount and structure of pure haplotype reconstruction error and
3. The combined influence of genotype error and haplotype reconstruction error on the overall haplotype error.

This work treats these aspects in three separate parts, building on each other consecutively and leading to a combined conclusion:

Characterization and quantification of genotyping error and its impact on genetic association studies:

So far, genotyping error models where either motivated by biological reasoning or mathematical simplicity. In most cases, a certain size of genotype error has just been assumed, often too high to be in a realistic range. Therefore, this investigation aims to close this gap. A representative sample of large epidemiological studies with duplicate genotypes has been collected to provide an approach to estimate misclassification probabilities in this routine data, and to characterize the model and the size of the genotyping error as it can be expected in genetic epidemiological practice. It was a further aim to elucidate the impact of such realistic genotype misclassification on genetic association data by applying the MC-SIMEX method.

Characterization and quantification of pure haplotype reconstruction error: A general understanding of the magnitude of the haplotype reconstruction error was to be achieved using a systematic approach by simulations and analytical derivations. Various simulation scenarios were applied including scenarios based on realistic haplotype distributions from epidemiological study data at hand. As researchers are usually interested in the associated risk of a specific haplotype, the error in this “risk haplotype” is of interest when interpreting the results. Therefore, this work focused on haplotype-specific error measures posing a classical misclassification problem. The sensitivity and specificity are presented as two intuitive measures for haplotype error, already known from classical misclassification problems. A systematic overview of haplotype error measures is given. Analytical and simulation approaches to quantify these error measures and to describe their size and dependencies are applied.

Haplotype misclassification from genotype error and haplotype reconstruction combined and its impact on association estimates: The final purpose of this work was to quantify the haplotype misclassification resulting from a combination of genotype error and haplotype reconstruction in individually assigned haplotypes and its impact on haplotype association estimates. Although several methods have been developed to incorporate haplotype uncertainty (see section 1.4.2.2), they lack flexibility in modeling and do not account for genotype error. Therefore, a simulation approach was set up to estimate haplotype misclassification matrices.

These misclassification matrices are necessary to calculate error-corrected association estimates using the MC-SIMEX method, which was applied here for the first time on genetic variables. The impact of haplotype misclassification as derived by the MC-SIMEX correction was exemplified on 15 SNPs of the *APM1* gene from 1770 subjects of the SAPHIR study [Heid et al., 2006] and on their association on plasma adiponectin concentrations.

Researchers that interpret the relevance of haplotype association estimates or use haplotype assignments to put up phylogenetic trees or to select certain persons for further studies, rely on sufficient haplotype certainty. Therefore, a thorough investigation of haplotype uncertainty, resulting from genotype error and reconstruction error, with special emphasis on misclassification probabilities of individually inferred haplotypes, is essential.

2. Quantification of genotyping error and its impact on genetic association studies

The error in SNP genotypes from an established method in an experienced lab is expected to be very small. This assumption has to be verified using data from routine genotyping as they would be used for association analysis. To estimate such small error, a representative set of large epidemiological studies with double SNP genotypes was collected. Misclassification probabilities for different error models were estimated in this data from routine genotyping using a Maximum Likelihood approach. Furthermore, the impact of such misclassification was elucidated on a real genetic association example.

2.1. Methods and material

2.1.1. The method of genotyping

Genotyping with the automated high-throughput MALDITOF-MS (Sequenom, San Diego) is processed by multiplying DNA-sections of a subject in a way that exactly the alleles at the locus of interest are duplicated manifold. The mass of these products is then measured via mass spectrometry. Figure 9 shows the mass spectrometry genotyping profile of alleles of one SNP for a heterozygous subject. The x-axis displays the mass of this product, the y-axis the frequency with which products of the respective mass are found. Each of the two alleles refers to a product with the specific mass. Therefore, a signal is detected either only at the first of the two positions at the x-axis (homozygous genotype of the first allele), only at the second of the two positions (homozygous genotype of the second allele), or at both (heterozygous genotype). The other signals in the picture are plain white noise. A software tests whether the amplitude of a signal exceeds a certain machine-determined level ("maximum detection level"). A "conservative" signal is detected, if the genotype was called unambiguously. If a genotype signal was detected only marginally or is even questionable, the output of the genotype detection software is said to be "moderate" or "aggressive".

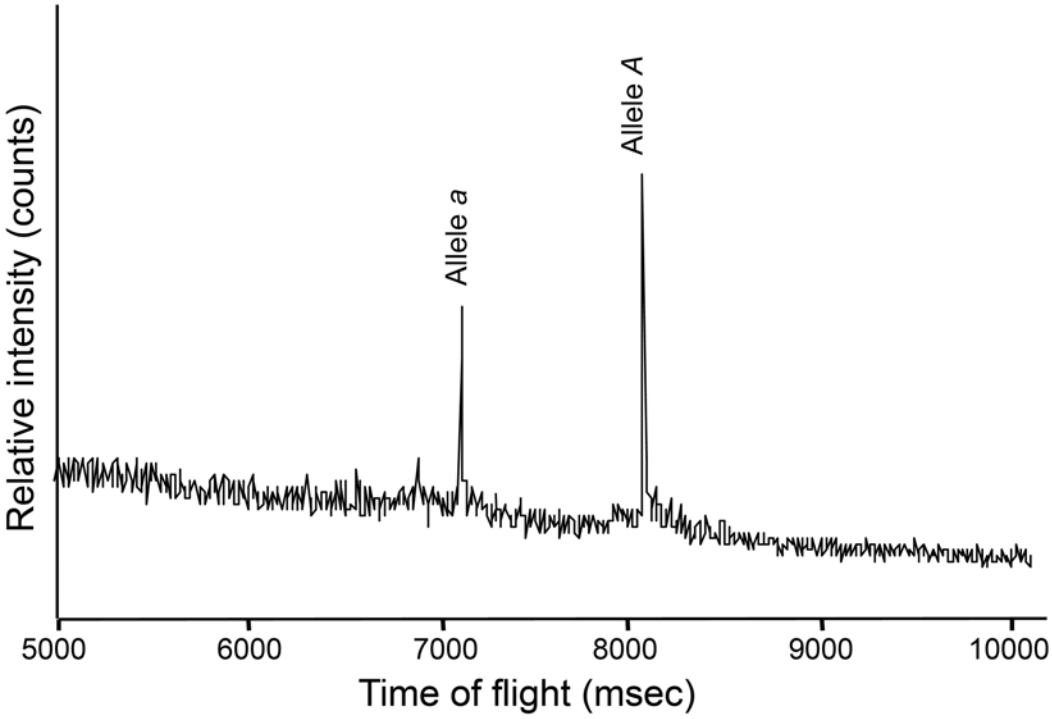


Figure 9: *Genotyping signal from MALDI-TOF MS: The x-axis displays the mass of the extension primer product and the y-axis the intensity of the product. Each of the two alleles refers to a product with specific mass. Therefore, a signal is detected at either of these two positions at the x-axis (homozygous genotype of either of the alleles) or at both (heterozygous as shown in the picture). The other signals are white noise.*

The algorithm's outcome can be verified by visual inspection, if a moderate or an aggressive call is detected, and changed, if necessary ("user call").

The white noise in the signal is unavoidable and one main reason of genotype error: If the true signal is not strong enough, it might vanish beneath the white noise. Then, this signal is not detected and the genotype is misclassified. For example, larger variability in the white noise in Figure 9 might cause an undetection of allele a and thus a misclassification of the heterozygous genotype aA in the homozygous genotype AA.

2.1.2. Collection of duplicate genotype data

In order to estimate the error in SNP genotypes from MALDI-TOF MS, information on genotypes were collected on a representative set of genetic epidemiological studies with two genotypes by person and SNP (*double genotypes*). The source of double genotypes was either routine quality control, that is, DNA of one subject was put on two positions of the same well

plate (*routine doubles*), or a well plate was run a second time due to insufficient call rate in the first run (*trouble-shooting doubles*).

All studies that were genotyped by the laboratory personnel of the GSF-Institute of Epidemiology in the years 2004-2005 by MALDI-TOF MS with about or more than 1000 subjects and 5% double genotypes were included. The final data set included eight studies with five distinct samples from the KORA studies [Wichmann et al., 2005], a study sample from Utah [Schoenborn et al., 2006], the SAPHIR study [Heid et al., 2006], and the German part of the AIRGENE study [Ruckerl et al., 2007].

Using these duplicate genotype data, the focus is set on the errors induced by genotyping process itself disregarding sample mismatch. Genotype error due to diluted or degraded DNA can only be covered partially.

2.1.3. Notation and definitions of genotypes

The total number of subjects is denoted with N and the total number of SNPs with L. The true genotype for SNP l of a subject i shall be denoted by G_{il}, the observed genotype by $G_{il}*$, and – if repeated measurements are available – by $G_{il}*^{(1)}$ and $G_{il}*^{(2)}$ for the first and second observed genotype. Observed genotype frequencies $P(G_l* = j)$, l=1,…,L, j=0,1,2, are denoted as $p*^{(l)} = (p*_0^{(l)}, p*_1^{(l)}, p*_2^{(l)})$, l=1,…,L with $p*_0^{(l)} = 1 - p*_2^{(l)} - p*_1^{(l)}$. True genotype probabilities shall be denoted by $\pi_j^{(l)} = P(X^{(l)} = j)$, j=0,1,2, $\pi^{(l)} = (\pi_0^{(l)}, \pi_1^{(l)}, \pi_2^{(l)})$ with $\pi_0^{(l)} = 1 - \pi_2^{(l)} - \pi_1^{(l)}$. Potentially error-prone observed genotype probabilities are denoted as $\pi*^{(l)} = (\pi*_0^{(l)}, \pi*_1^{(l)}, \pi*_2^{(l)})$ with $\pi_0*^{(l)} = 1 - \pi_1*^{(l)} - \pi_2*^{(l)}$, for which $p*^{(l)}$ is a consistent estimate.

2.1.4. Discordance matrix

For each SNP l, l=1,...L, the number of concordant and discordant observed genotype pairs are derived by $R^{(l)} = (r_{ij}^{(l)})_{i,j=0,1,2}$ with $r_{ij}^{(l)}$ being the number of subjects with $G_l*^{(1)} = i$ and $G_l*^{(2)} = j$ (*discordance matrix*). Summing the $r_{ij}^{(l)}$ over i and j yields the total number of observed genotype pairs, for each l=1,...L giving rise to the restriction $\sum_{i,j} r_{ij}^{(l)} = N$. Without

order of the measurements, it is irrelevant which measurement is denoted as $G_l{}^{*(1)}$ or $G_l{}^{*(2)}$, and thus the matrix is triangular as in Table 1. The *overall discordance* was computed as the number of discordant pairs across all SNPs relative to the total number of double genotype pairs. The *SNP-wise discordance* was computed as the number of discordants divided by the number of double genotype pairs per SNP.

Table 1: *Triangular discordance matrix for the lth SNP*

		$G_l{}^{*(2)}$		
		0	1	2
$G_l{}^{*(1)}$	0	$r_{00}{}^{(l)}$		
	1	$r_{01}{}^{(l)}$	$r_{11}{}^{(l)}$	
	2	$r_{02}{}^{(l)}$	$r_{12}{}^{(l)}$	$r_{22}{}^{(l)}$

2.1.5. Misclassification matrix and the problem of identifiability

Based on the discordance matrix, the misclassification probabilities were estimated, i.e. the probabilities that a true genotype $G=j$ was misclassified as $G^*=i$, $i,j=0,1,2$, constituting the misclassification matrix, $\Pi^{(l)} = (\pi_{ij})^{(l)}_{i,j=0,1,2} = (P(G^* = i \mid G = j))^{(l)}_{i,j=0,1,2}$, with $\pi^{(l)}_{0j} = 1 - \pi^{(l)}_{1j} - \pi^{(l)}_{2j}$.

Therefore, the misclassification problem in the SNP genotype is represented by a 3x3 misclassification probability matrix for each SNP with each cell describing the probability that a true genotype 0,1, or 2 is measured as 0,1, or 2 (using the number of minor alleles as genotype coding). Thus, there are nine misclassification probabilities to estimate. Since each column sums up to unity due to the restriction $\pi^{(l)}_{0j} = 1 - \pi^{(l)}_{1j} - \pi^{(l)}_{2j}$, there are only six unknown parameters in each general misclassification matrix.

Such a misclassification matrix is the heart of any misclassification problem defining the model as well as the size of the error, and it is the prerequisite for statistical methods accounting for the misclassification in association analysis. The methodology to solve this misclassification problem in general requires more than two measurements for validation data, but repeated genotyping for routine quality control is not usually performed more than

twice. Thus, statistical procedures requiring more than two measurements cannot be applied [Lai et al., 2007;Pompanon et al., 2005]. This leaves the problem of making this 3x3 misclassification problem identifiable with double measurements. "Not identifiable" means that there are more parameters to estimate than there is information to rely upon: In case of L SNPs and 6 unknown parameters for each SNP's misclassification matrix, there are $6L$ parameters to estimate from $5L$ independent observations, which are the number of subjects with genotype i in the first and genotype j in the second measurement for the lth SNP, $r_{ij}^{(l)}$, with $\sum_{i,j} r_{ij}^{(l)} = N$ for all l=1,…,L, i,j=0,1,2, and N being the number of subjects. Identifiability was achieved by assuming equal misclassification matrices for all SNPs.

The general 3x3 misclassification matrix $\Pi = (\pi_{ij})_{i,j=0,1,2} = (P(G^* = i \mid G = j))_{i,j=0,1,2}$ on the three-level genotype (i.e. SNPs with non-missing minor allele homozygote category) is illustrated in Table 2. The three constraints $1 = \sum_{j=0,1,2} \pi_{ij}$, for i=0,1,2 leave 6 unknown parameters $(\pi_{01}, \pi_{02}, \pi_{10}, \pi_{12}, \pi_{20}, \pi_{21})$.

Table 2: *Notation of a general genotype misclassification matrix (unrestricted error model)*

		True genotype G		
		0	1	2
Observed genotype G^*	0	π_{00}	π_{01}	π_{02}
	1	π_{10}	π_{11}	π_{12}
	2	π_{20}	π_{21}	π_{22}

with $\pi_{00} = 1 - \pi_{10} - \pi_{20}$, $\pi_{01} = 1 - \pi_{11} - \pi_{21}$, $\pi_{02} = 1 - \pi_{12} - \pi_{22}$

2.1.6. Genotype error models

To estimate misclassification matrices, error models have to be assumed. Based on the genotyping mechanism explained in section 1.4.1 and due to the underlying white noise measured (see section 1.4.1, Figure 9), several error models are discussed, which involve restrictions on parameters:

(A) <u>*Allelic drop out model*</u>: A true signal falls short of the detection level resulting in allelic drop-out, which implies that (i) a heterozygous subject is more likely misclassified as homozygous (one of the two signals vanished) than the other way

round, (ii) a subject homozygous for the one allele is unlikely misclassified as homozygous for the other allele, and (iii) a homozygous subject is more likely coded as missing than a heterozygous subject [Mitchell et al., 2003]. This model corresponds to the following restrictions: $\pi_{01} > \pi_{10}$ and $\pi_{21} > \pi_{12}$, involving 6 parameters

(B) The *zero-corner model* [Morris and Kaplan, 2004], which reflects an extreme case of (A)(ii) assuming a zero probability of a homozygous genotype being misclassified as the other homozygous genotype, leading to the restrictions $\pi_{20} = 0$ and $\pi_{02} = 0$ reducing to 4 parameters

(C) The *symmetrical model* assuming no systematic ordering of the major and minor allele on the assay, which implies that misclassifying a true homozygous genotype or for falsely classifying as a homozygous genotype does not depend upon whether this is the minor or major allele. The following restrictions apply: $\pi_{10} = \pi_{12}$, $\pi_{01} = \pi_{21}$ and $\pi_{20} = \pi_{02}$ reducing to 3 parameters

(D) The *allele-independent model*, which assumes that the probability of misclassifying one allele into the other is the same as the other way round [Akey et al., 2001b] , corresponding to *P*(Allele A is misclassified into Allele a) = *P*(Allele a is misclassified into Allele A) =: ε , reducing to 1 parameter. The misclassification matrix for this most parsimonious model is illustrated in Table 3.

Table 3: *Misclassification matrix for allele-independent model*

		True genotype *G*		
		0	1	2
Observed genotype *G**	0	$1-2\varepsilon(1-\varepsilon)-\varepsilon^2$	$\varepsilon(1-\varepsilon)$	ε^2
	1	$2\varepsilon(1-\varepsilon)$	$1-2\varepsilon(1-\varepsilon)$	$2\varepsilon(1-\varepsilon)$
	2	ε^2	$\varepsilon(1-\varepsilon)$	$1-2\varepsilon(1-\varepsilon)-\varepsilon^2$

The other genotyping error models described in the literature are closely related to the four already stated above with an exception being the “uniform error model” [Lincoln and Lander, 1992], which assumes equal misclassification probabilities. This model is mathematically appealing but rather unrealistic for this genotype setting.

2.1.7. Estimating the genotype misclassification via maximum-likelihood

The observed discordances $r_{ij}^{(l)}$ (Table 1) are estimates for the discordance probabilities, i.e. the probabilities of observing genotype *i* at the first measurement and genotype *j* at the second with arbitrary ordering of the measurements, $\delta_{ij}^{(l)} = \Pr ob(G*^{(1)} = i \wedge G*^{(2)} = j)$ for l=1,…L, i,j=0,1,2, $i<j$. These relate to the misclassification probabilities $(\pi_{ij})_{i,j=0,1,2}$ and the true genotype probabilities $(\pi^{(l)})_{l=1,\ldots,L}$ via

$$\begin{aligned}
\delta_{00}^{(l)} &= \pi_2^{(l)}\pi_{02}{}^2 + \pi_1^{(l)}\pi_{01}{}^2 + \pi_0^{(l)}\pi_{00}{}^2 \\
\delta_{11}^{(l)} &= \pi_2^{(l)}\pi_{12}{}^2 + \pi_1^{(l)}\pi_{11}{}^2 + \pi_0^{(l)}\pi_{10}{}^2 \\
\delta_{22}^{(l)} &= \pi_2^{(l)}\pi_{22}{}^2 + \pi_1^{(l)}\pi_{21}{}^2 + \pi_0^{(l)}\pi_{20}{}^2 \\
\delta_{02}^{(l)} &= 2\pi_2^{(l)}\pi_{02}\pi_{22} + 2\pi_1^{(l)}\pi_{01}\pi_{21} + 2\pi_0^{(l)}\pi_{00}\pi_{20} \\
\delta_{01}^{(l)} &= 2\pi_2^{(l)}\pi_{02}\pi_{12} + 2\pi_1^{(l)}\pi_{01}\pi_{11} + 2\pi_0^{(l)}\pi_{00}\pi_{10} \\
\delta_{12}^{(l)} &= 2\pi_2^{(l)}\pi_{12}\pi_{22} + 2\pi_1^{(l)}\pi_{11}\pi_{21} + 2\pi_0^{(l)}\pi_{10}\pi_{20}
\end{aligned} \qquad (1)$$

When the true genotype frequencies $\pi^{(l)}$ are known, the likelihood for $R^{(l)}$ given Π, $L_{R,p*}(\Pi) := L(\Pi \mid (R^{(l)}, p*^{(l)})_{l=1,\ldots,L})$, is given as

$$\prod_l (\delta_{00}^{(l)})^{r00(l)} (\delta_{11}^{(l)})^{r11(l)} (\delta_{22}^{(l)})^{r22(l)} (\delta_{02}^{(l)})^{r02(l)} (\delta_{01}^{(l)})^{r01(l)} (\delta_{12}^{(l)})^{r12(l)} \qquad (2).$$

This likelihood can be interpreted as the probability that a certain misclassification matrix Π is true given the observed discordance matrices and genotype frequencies. The misclassification matrix with the highest probability fits best to the observed data. Therefore, misclassification probabilities are estimated by maximizing this likelihood (Maximum Likelihood approach, ML).

When the true genotype probabilities $\pi^{(l)}$ are unknown, they either need to be estimated together with the misclassification probabilities ("extended likelihood") or assumptions need to be made. Applying the latter approach, the following assumptions were made: (a) the observed genotype probabilities reasonably approximate the truth ($\pi^{(l)} \approx \pi*^{(l)}$) and (b) $\pi*^{(l)}$ is estimated by $p*^{(l)}$ with negligible sampling error. Note that $p*^{(l)}$ is the maximum likelihood (ML) estimate of $\pi*^{(l)}$ for the likelihood $\prod_l \pi*_2^{(l)}{}^{np*_2^{(l)}} \; \pi*_1^{(l)}{}^{np*_1^{(l)}} \; \pi*_0^{(l)}{}^{np*_0^{(l)}}$.

These two assumptions result together in $\pi^{(l)} \approx p*^{(l)}$ (*small misclassification / large sample assumption*).

Therefore, misclassification probabilities, Π, were estimated by maximizing $L_{R,p^*}(\Pi)$ as given in equation (2) with $\pi^{(l)} \approx p^{*(l)}$. Again assuming small misclassification, exact p-values were computed to test for Hardy-Weinberg equilibrium (HWE) using $\pi^{*(l)}$.

In sensitivity analyses, the robustness of the estimation was evaluated upon violation of the assumption $\pi^{(l)} \approx p^{*(l)}$, upon exclusion of SNPs with HWE violation, or upon exclusion of SNPs with only few subjects homozygous of the minor allele. Finally, it was explored what the results would have looked like, if the *small misclassification assumption* had not been chosen, but if the true genotype probabilities had been estimated simultaneously with the misclassification probabilities (2*K*+6 parameter to estimate). Thus, the assumption $\pi^{(l)} \approx p^{*(l)}$ was abandoned for this sensitivity analysis and the extended likelihood given by the product of equation (2) and $\prod_l \pi^{*(l)}_2{}^{np^{*(k)}_2} \; \pi^{*(l)}_1{}^{np^{*(k)}_1} \; \pi^{*(l)}_0{}^{np^{*(k)}_0}$ was maximized utilizing the relationships $\pi^{*}_0{}^{(l)} = 1 - \pi^{*}_1{}^{(l)} - \pi^{*}_2{}^{(l)}$, $\pi^{*}_1{}^{(l)} = \pi_2^{(l)} \cdot \pi_{12} + \pi_1^{(l)} \cdot \pi_{11} + \pi_0^{(l)} \cdot \pi_{10}$, and $\pi^{*}_2{}^{(l)} = \pi_2^{(l)} \cdot \pi_{22} + \pi_1^{(l)} \cdot \pi_{21} + \pi_0^{(l)} \pi_{20}$.
All maximum likelihood (ML) estimates were computed by applying the Nelder-Mead simplex algorithm implemented in Mathematica (version 5.0, Wolfram Research, Champaign, IL). The variance of the ML estimator was derived by the Fisher matrix, which is the value of the second derivation of the likelihood at the ML estimator value. The misclassification was estimated based on the general as well as on the restricted models (see A-D, section 2.1.6). Furthermore, likelihood ratio tests were conducted to compare model-fit.

2.1.8. Correction of association from *APM1* SNPs on Adiponectin

The association of the *APM1* gene SNPs with adiponectin plasma levels, which is described in section 1.2.6, was re-analyzed to elucidate the impact of the SNP genotype misclassification on association analysis. Since homozygote rare genotypes were missing in 2 of the 15 SNPs, the analysis was restricted to 13 SNPs with all possible genotype values present. For each of these SNPs, a linear regression model on log(adiponectin+1) adjusted for BMI, sex and age was applied. The estimated regression coefficients without correcting for misclassification (naive estimates) were contrasted with the corrected MC-SIMEX estimates applying the loglinear extrapolation function assuming two different dimensions of misclassifications: a realistic scenario based on the general misclassification matrix estimated

from the double genotypes, and an extreme scenario yielded by multiplying the non-diagonal elements of the realistic scenario matrix by 10.

2.2. Results

2.2.1. Description of duplicate genotype data

The data set contained 646558 genotypes with 160454 doubles involving 283 SNPs from over 10000 subjects in 8 projects. Among these were 70539 routine doubles. For 62318 routine doubles, both genotype measurements were non-missing; 57805 of these corresponded to 225 "three-level" SNPs, which are SNPs where the number of subjects homozygous of the minor allele was non-zero and thus all three genotype categories existed. Table 4 summarizes the sample sizes and number of duplicates for each project.

Table 4: *Summary of collected data from eight included projects*

Project number	1	2	3	4	5	6	7	8	∑
# of subjects	1080	930	1830	2489	1628	1776	2907	1400	14040
# of SNPs per subject	98-115	37-46	15-19	13-17	15	9-18	34-44	1-9	283
# of genotypes	258517	79646	36480	41999	36483	41216	130383	21834	646558
# of measured Person-SNPs*	114464	37695	33590	40085	24426	25794	112003	10917	398974
# of **all** doubles (% of Person-SNPs)	90162 (78.77)	17732 (47.04)	2787 (8.30)	1914 (4.77)	5696 (23.32)	14260 (55.28)	16986 (15.17)	10917 (100)	160454 (40.22)
# of SNPs with routine doubles§	111	37	19	17	14	18	37	9	262
# of **routine** doubles (% of Person-SNPs)	33262 (29.06)	4347 (11.53)	2787 (8.30)	1914 (4.77)	2960 (12.12)	590 (2.29)	13762 (12.29)	10917 (100)	70539 (17.68)
# of **routine** doubles (% on Person-SNPs) without missings and 3 genotype values present	26832 (23.44)	3199 (8.49)	2664 (7.93)	1188 (2.96)	1980 (8.11)	241 (0.93)	11957 (10.68)	9744 (89.26)	57805 (14.49)

* "*Person-SNPs*" = the product of number of persons times number of SNPs. § "doubles" = double genotypes, i.e. pairs of two genotype measurements on the same subject and same SNP.

2.2.2. Discordance between duplicate genotypes

The discordance matrix including all doubles - i.e. routine as well as trouble-shooting doubles - summarizing over all 283 SNPs is shown in Table 5a. This matrix depicts the missing values as one genotype category. This table highlights that the proportion of missings among the first measurements is 15.65% versus 8.15% among the second (shaded grey in the table) indicating an undesirable informative ordering of the measurements. Restricting to the routine doubles, now including 262 SNPs, leads to symmetric proportions of missing genotypes (7.56% versus 7.60%, respectively, Table 5b) indicating that missingness is now independent of the measurement order. The main analysis was based on the 225 three-level SNPs with 57805 routine double genotypes, both non-missing, which yielded 210 discordant pairs and thus an overall discordance of 0.36%. Table 6 depicts the discordance across all SNPs as a triangular matrix.

Table 5: *Observed duplicate genotypes from measurement 1 ($G*^{(1)}$) and 2 ($G*^{(2)}$) including missing values as category.*

(a) *routine and trouble-shooting genotype doubles.*

Freq Percent Row Col Perc.		$G*^{(2)}$ 0	1	2	Miss	∑
$G*^{(1)}$	0	76181 47.48 94.88 87.12	156 0.10 0.19 0.32	18 0.01 0.02 0.15	3934 2.45 4.90 30.06	80289 50.04
	1	80 0.05 0.18 0.09	41774 26.03 95.19 86.63	43 0.03 0.10 0.37	1989 1.24 4.53 15.20	43886 27.35
	2	19 0.01 0.17 0.02	160 0.10 1.43 0.33	10421 6.49 93.34 89.03	564 0.35 5.05 4.31	11164 6.96
	Miss	11162 6.96 44.44 12.77	6132 3.82 24.42 12.72	1223 0.76 4.87 10.45	6598 4.11 26.27 50.42	25115 15.65
	∑	87442 54.50	48222 30.05	11705 7.29	13085 8.15	160454 100.00

(b): *routine genotype doubles only*

Freq Percent Row Col Perc.		$G*^{(2)}$ 0	1	2	Miss	∑
$G*^{(1)}$	0	36202 51.32 95.53 95.53	70 0.10 0.18 0.32	10 0.01 0.03 0.18	1613 2.29 4.26 30.09	37895 53.72
	1	55 0.08 0.25 0.15	20746 29.41 95.24 95.19	30 0.04 0.14 0.55	952 1.35 4.37 17.76	21783 30.88
	2	10 0.01 0.18 0.03	37 0.05 0.67 0.17	5158 7.31 93.36 93.97	320 0.45 5.79 5.97	5525 7.83
	Miss	1628 2.31 30.51 4.30	941 1.33 17.63 4.32	291 0.41 5.45 5.30	2476 3.51 46.40 46.19	5336 7.56
	∑	37895 53.72	21794 30.90	5489 7.78	5361 7.60	70539 100.00

Table 6: *Observed triangular discordance matrix: restricted to routine doubles (measurement 1 ($G*^{(1)}$) and 2 ($G*^{(2)}$)) with both measurements non-missing and three-level SNPs (225 SNPs, 57805 double genotypes).*

Frequency Percent		$G*^{(2)}$ 0	1	2
$G*^{(1)}$	0	32498 56.2201		
	1	123 0.2128	19944 34.5022	
	2	20 0.0346	67 0.1159	5153 8.9145

To illustrate potential dependencies of the discordance on HWE violation or minor allele frequency (MAF), scatter plots were drawn of the SNP-wise discordance versus p-value from testing for HWE violation (Figure 10a) or versus MAF (Figure 10b). It can be seen that some of the larger discordances occurred together with smaller HWE p-values, but not all HWE violations implicated large discordance (Spearman correlation coefficient r=-0.1362, p=0.0313). There was no dependency of the discordance on the MAF (r= 0.0826, p=0.1927).

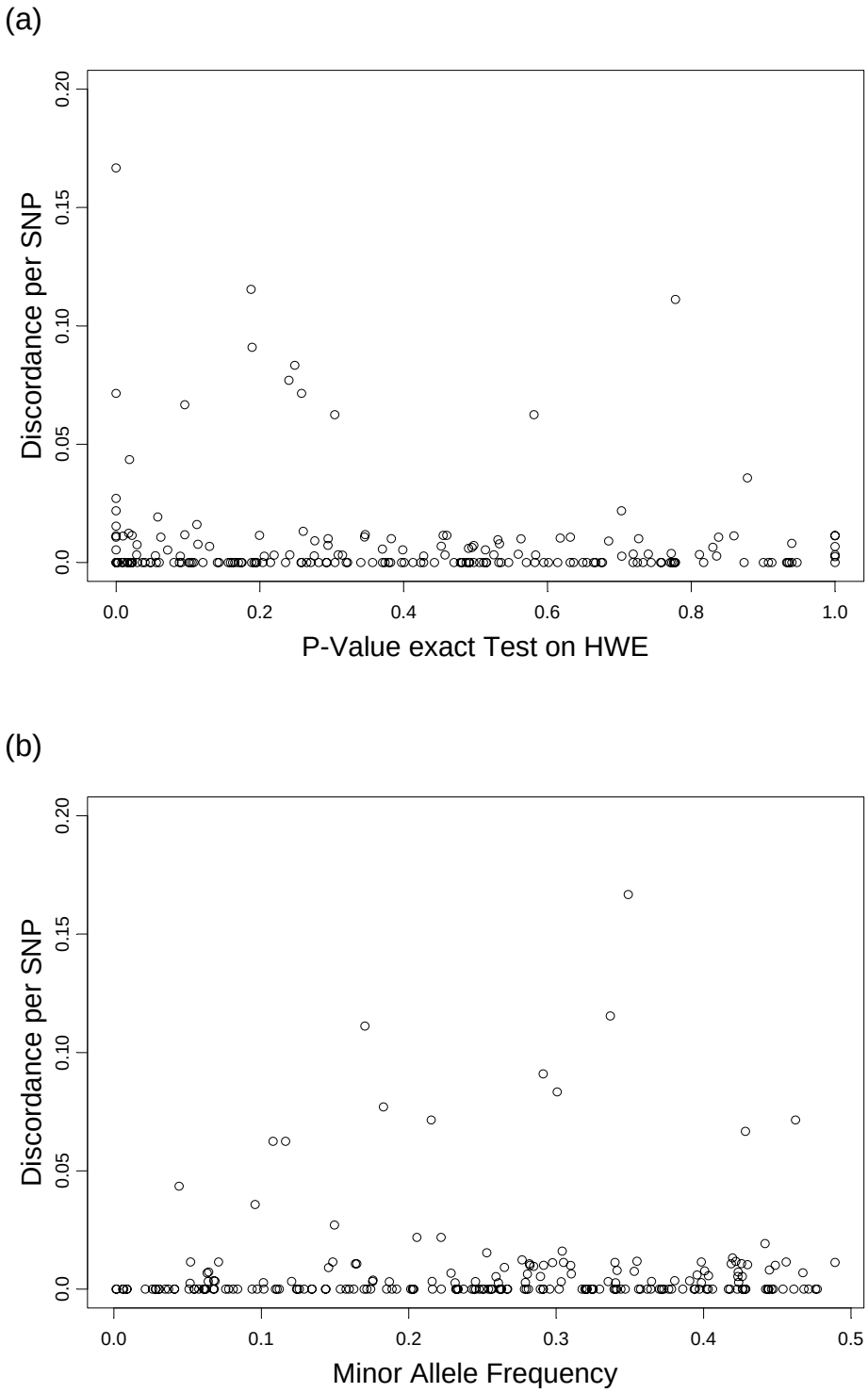


Figure 10: *Scatter plot of the SNP-wise discordance observed in the 57805 double genotypes with both measurements non-missing on 225 three-level SNPs versus (a) the p-value from testing for violation of Hardy-Weinberg-Equilibrium (HWE) and (b) the minor allele frequency.*

2.2.3. Estimated genotype misclassification matrices

Quantification of misclassification matrix

Table 7 summarizes the misclassification matrices from maximizing the likelihood $L_{R,p^*}(\Pi)$ for the various misclassification models.

Table 7: *Estimated misclassification matrix under various misclassification models, showing the estimate and 95% confidence intervals.*

(a) *general misclassification model (6 parameter, unrestricted)*

		true genotype *G*		
		0	1	2
observed genotype *G**	0	0.999505	0.0024277 [0.0016900, 0.0031655]	0.0013803 [0.0004305, 0.0023300]*
	1	0.000391 [0.000014, 0.000768]	0.996023	0.000229 [-0.000450, 0.000907]
	2	0.000104 [-0.000050, 0.000258]	0.001549 [0.001034, 0.002065]	0.9983911

(b) *assuming zero corner model (4 parameter)*

		true genotype *G*		
		0	1	2
observed genotype *G**	0	0.999880	0.003465 [0.002720, 0.004210]	0**
	1	0.000120 [-0.000179,0.000419]	0.995136	0.002979 [0.001207,0.004751]**
	2	0	0.001399 [0.000880, 0.001917]	0.997021

(c) *assuming symmetric model (3 parameter)*

		true genotype *G*		
		0	1	2
observed genotype *G**	0	0.998740	0.001436 [0.000911, 0.001961]	0.000264 [0.000148,0.000380]**
	1	0.000997 [0.000257,0.001736]	0.997127	0.000996 [0.000257,0.001736]
	2	0.000264 [0.000147, 0.000380]	0.001436 [0.000912, 0.001961]	0.998740

(d) *assuming allele-independent model (1 parameter)*

		true genotype *G*		
		0	1	2
Observed genotype *G**	0	0.998008	0.000996 [0.000867,0.001124]**	0.0000009 [0.0000007,0.0000013]**
	1	0.001991 [0.001734,0.002249]**	0.998009	0.001991 [0.001734,0.002249]**
	2	0.0000009 [0.0000007,0.0000013]**	0.000996 [0.000867,0.001124]**	0.998008

* does not include zero; thus zero-corner model is not supported.

** no overlap with 95% CI of general model

For the general error model, the estimated misclassification probabilities ranged between 0.000104 to 0.0024277 (Table 7a). The zero corner model and the symmetrical model yielded slightly different result, but very close regarding the dimension of the error (Table 7b,c). Assuming the allele-independent error model, an ε-parameter of 0.000997 was estimated resulting in misclassification probabilities between 0.0000009 and 0.001991 (see Table 7d).

Characterization of misclassification model

To evaluate the misclassification model most supported by the duplicate data, 95% confidence intervals of parameters and likelihood ratio tests were conducted. The estimated parameters and 95% confidence intervals indicate that the allelic-drop out characteristic holds ($\pi_{10} < \pi_{01}$ and $\pi_{12} < \pi_{21}$), but that the restricted models (zero-corner, symmetrical, allele-independent) did not completely fit the data. The symmetric model deviated the least from the general model as the 95% confidence interval from only one misclassification probability was disjoint with the corresponding general error model interval. This is illustrated in Figure *11*.

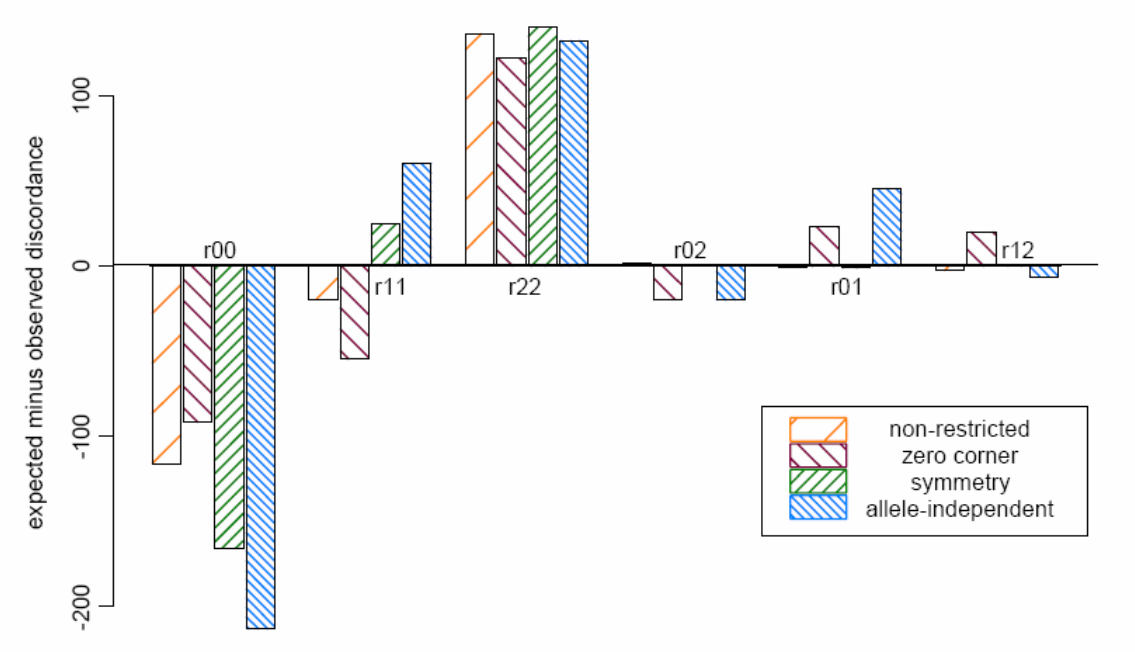


Figure 11: *Difference between observed and expected number of discordant pairs under the various error models as a measure of model-fit. The r_{ij} denotes the number of subjects with one measurement yielding genotype i and the other genotype j, for i,j=0,1,2, i<j. The general model (unrestricted) provided the best fit.*

The observed number of discordants with one major allele homozygous and the other minor allele homozygous, r_{02}, was higher than expected, when π_{20} and π_{02} were assumed to be zero (zero-corner model) or close to zero (allele-independent model). The observed number of discordants with one heterozygous and one major allele homozygous, r_{01}, was higher than expected when assuming symmetry (symmetric or the allele-independent model).

The likelihood ratio test on model fit (Table 8) comparing to the unrestricted model, yielded no formal rejection of the symmetrical model (though a "borderline" p-value of 0.07), but for the zero corner and the allele-independent model ($p< 10^{-3}$).

Table 8: *Likelihood ratio test results for Goodness-of-Model-Fit: Comparing the restricted models A-D with the general error model, also stating the number of observed genotype pairs and the number of discordant genotype pairs as expected under the various error models*

Model	# of discordant genotype pairs r_{02}, r_{01}, r_{12}	# of parameter	Log-likelihood	λ comparing with general model (df)	p-value
observed	20, 123, 67				
General model (A)	21.5, 122.5, 64.4	6	-46768.5	-	-
Zero corner model (B)	0.2, 146.3, 87.4	4	-46834.1	131.2 (2)	$<10^{-3}$
Symmetric model (C)	19.9, 122.1, 68.1	3	-46772.1	7.2 (3)	0.07
Allele-independent model (D)	0.1, 168.7,61.0	1	-46861	185 (5)	$<10^{-3}$

with $\lambda = -2*(\ln L_{restricted} - \ln L_{general}) \sim \chi^2_{df}$; df= 6 - # parameters of restricted model; the r_{02}, r_{01}, r_{12} referring to the notation in Table 1.

Robustness of estimation

In the sensitivity analysis, the impact of violation of the small misclassification assumption was explored, i.e. deviation of $\pi^{(l)}$ from $p^{*(l)}$, l=1,…, 225. The deviation was chosen such that $(p^{*(l)})_{1,\ldots,225}$ would have been observed given an allele-independent error with $\varepsilon = 0.001$. The new $\pi^{(l)}$ were thus derived from $\pi^{*}{}_0^{(l)} = 1 - \pi^{*}{}_1^{(l)} - \pi^{*}{}_2^{(l)}$, $\pi^{*}{}_2^{(l)} = (1-2\varepsilon)\pi_2^{(l)} + (\varepsilon - 2\varepsilon^2)\pi_1^{(l)} + \varepsilon^2$, and $\pi^{*}{}_1^{(l)} = (1-4\varepsilon+4\varepsilon^2)\pi_1^{(l)} + 2\varepsilon - 2\varepsilon^2$. The misclassification probabilities were again estimated maximizing $L_R(\Pi)$ given the new $\pi^{(l)}$. For the general error model, the estimated parameters $(\pi_{01}, \pi_{02}, \pi_{10}, \pi_{12}, \pi_{20}, \pi_{21})$ were similar with (0.0024, 0.0016, 0.0004, 0.0002, 0.0000, 0.0016) instead of (0.0024, 0.0014, 0.0004,

0.0002, 0.0001, 0.0015). The ε-parameter for the allele-independent model remained basically unchanged with 0.0010.

When excluding the 29 SNP with p-values < 0.05 for the test of HWE violation, the results did not change markedly.

When excluding the SNPs with less than 30 subjects in the homozygous of the minor allele genotype category (leaving 152 SNPs), the estimated general model misclassification probabilities were (0.0018, 0.0013, 0.0009, 0.0002, 0.0002, 0.0018) and the ε-estimate was 0.0012.

Finally, when estimating the true genotype probabilities together with the misclassification probabilities, the analysis had to be restricted to the 152 SNPs with enough observations in the third genotype category, resulting in the estimates (0.0020, 0.0011, 0.0008, 0.0061, 0.0002, 0.0000). The ε-estimate remained unchanged when compared to the previous analysis with 152 SNPs at 0.0012.

In summary, sensitivity analyses yielded negligable deviations for estimation of the ε-parameter for the allele-independent model, and slightly different results for the general error model. Altogether, misclassification probabilities remained very small throughout.

2.2.4. *APM1* data example: Corrected genotype association estimates

Figure 12 summarizes the uncorrected (naive) and the MC-SIMEX corrected β-estimates in the example of the 13 three-level *APM1* SNPs and their association with plasma adiponectin concentrations. A clear change of the β-estimates of up to 15% when correcting for the misclassification was only seen for two SNPs under the extreme scenario with already high naive estimates (SNP4 and SNP13).

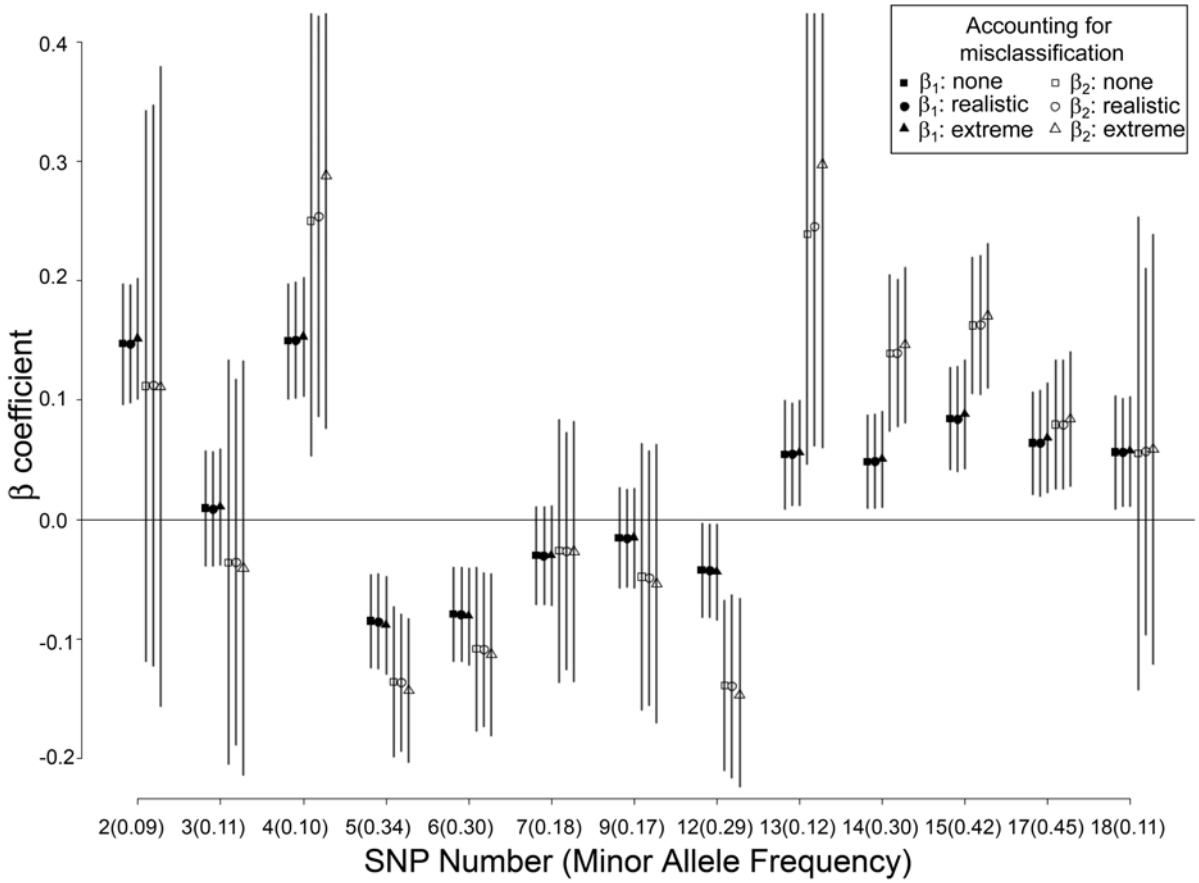


Figure 12: *Impact of Genotyping Error on SNP Association: The β coefficient estimates and 95% confidence intervals of the association of 13 APM1 SNPs with plasma adiponectin levels computed by linear regression adjusted for age, sex, and BMI are shown without accounting for misclassification (square), with correcting for a realistic (circle), and for extreme misclassification (triangle). The assumed realistic misclassification was the general error model as in Table 7a; the extreme scenario was using ten-fold as large misclassification probabilities (for non-diagonal elements). The β coefficient describes the unit increase in log(adiponectin+1) comparing the heterozygous (β_1) or the homozygous of the minor allele (β_2) with the homozygous of the major allele. SNP numbering refers to original publication [Heid et al., 2006].*

3. Quantification of haplotype reconstruction error

In the following chapter, the size and structure of haplotype reconstruction error is approximated. To focus on the error induced by the haplotype reconstruction alone, no genotype error is assumed. An overview of haplotype error measures is given, proposing sensitivity and specificity as error measures. Analytical and simulation approaches with various simulation scenarios are applied to quantify these error measures and to describe their size and dependencies.

3.1. Methods and Material

3.1.1. Notation and Definitions of Haplotypes

Notation of true haplotypes

Let L be the *number of loci* and N the *number of individuals*. For each individual i=1, ...,N, the vector $G_i = (G_{i1},...,G_{iL})$ denotes the subject's *genotypes* at the L loci, l=1, …L, with G_{il} indicating the number of minor alleles at locus l for individual i and $G_{il} \in \{0,1,2\}$. Consequently, there are 3^L possible values $\gamma=(\gamma_1,..., \gamma_L)$ for G_i. There are $M=2^L$ possible different haplotypes h_1, …, h_M in the population. The haplotypes of subject i can be written as a vector $H_i = (H_{i1},…,H_{iM})$, with each H_{im} indicating *the true number of copies of the haplotype* h_m of subject i, m=1, ..., M, and $H_{im} \in \{0,1,2\}$. Due to the restriction $\sum_{m=1,...,M} H_{im} = 2$, there are $\binom{M+1}{2}$ possible values $\eta=(\eta_1,..., \eta_M)$ for H_i. H_i thus denotes the individual's haplotype pair (*diplotype*) and the various η reflect all possible pairs. The number of different pairs actually appearing in a sample is further restricted by the correlation between the alleles at the loci. The *effective number of loci*, L_{eff}, can be computed according to Nyholt [Nyholt, 2004] taking this correlation into account.

Notation of reconstructed haplotypes

When statistically reconstructing haplotypes from genotypes, the *reconstructed number of copies of each haplotype* in subject i is denoted as $H_i^*=(H_{i1}^*,…,H_{iM}^*)$ being the vector of the expected values given the observed genotypes G_i as estimated by a reconstruction program: $H_i^*=E(H_i|G_i)$. As an unambiguous decision for a haplotype pair is not always possible, the

H_{im}* move in a continuous space, H_{im}* $\in$ [0,2]. The *most probable or most likely number of haplotypes* C_{im}* is derived by categorizing H_{im}* into the most likely haplotype pair for each individual with C_{im}* indicating *the observed number of copies of the haplotype* h_m, thus returning to the discrete space, C_{im}* $\in \{0,1,2\}$. They are also often denoted as individually inferred haplotypes.

Haplotypes can be inferred unambiguously (i.e. without error) for subjects being heterozygous in less than two loci. The *ambiguity fraction* is the number of subjects being heterozygous for at least two loci, N_{amb}, divided by N, which thus describes the proportion of the sample where haplotype reconstruction error might occur.

The *frequencies* of the haplotypes $h_1,\ldots, h_M$ in the sample are denoted as $f=(f_1,\ldots, f_M)$ with $f_m = \sum_{i=1,\ldots,N} H_{im} / 2N$, $m=1, \ldots, M$, $1 = \sum_{m=1,\ldots,M} f_m$. The sampling error for estimating the frequency is considered to be ignorable in large enough data sets. Analogously, the *frequencies of the reconstructed haplotypes* H_i* are denoted as f*$=(f_1$*$,\ldots, f_M$ *).

3.1.2. Haplotype reconstruction methods

The methods commonly used for reconstructing haplotypes are based primarily on the EM-algorithm (e.g. implemented in PROC HAPLOTYPE from SAS/GENETICS [Czika et al. 2002] or haplo.em in the R library haplo.stats) or on a Bayesian approach, like the PHASE algorithm [Stephens and Donnelly, 2003]. The latter method approximates the posterior distribution of haplotype probabilities using prior information based on the coalescent theory [Kingman, 1982]. This population genetic theory predicts the distribution of haplotypes in natural populations assuming one ancestor haplotype incorporating recombination and mutation. Unresolved haplotypes are therefore assumed to be similar to known or already resolved haplotypes und chosen to follow a certain probability distribution. Advantages of this method are only expected, if real observed data follow this theoretical concept.

Due to its relative ease of computation, methods based on the EM algorithm are often preferred. In general, the EM is an iterative process, which first calculates the expected likelihood of unknown parameters based on suitably chosen starting values given the observed data (Expectation-Step). In the Maximization Step, the parameters are re-estimated by maximization of the expected likelihood. This process is repeated, till convergence is reached. Relating to haplotypes, haplotype probabilities are chosen, which maximize the probability of the observed data. This method is briefly described here:

The probability of genotype G_i of the ith individual is given by $P(G_i|f_1,...,f_M) = \sum_{\eta\sim\gamma} \hat{f}_\eta *$, where $\eta\sim\gamma$ denotes the set of η's being consistent with the genotype γ and $\hat{f}_\eta *$ being the estimated probability of a subject having haplotype pair η. $\hat{f}_\eta *$ is given as $2\hat{f}_j * \hat{f}_k *$ for $j \neq k$ or $(\hat{f}_j *)^2$ otherwise, when η depicts the haplotype pair h_j / h_k, j,k=1,…,M. Assuming that the genotypes are independently distributed, the log-likelihood of the sample is

$$\log L = \sum_{i=1}^{N} \log P(G_i|f_1,...,f_M) .$$

The EM-algorithm is an iterative procedure beginning with the random assignment of initial values $\hat{f}_1 *^{(0)},...,\hat{f}_M *^{(0)}$ for the population haplotype frequencies $f_1,…,f_M$. In the E-step of the procedure, the above probability $P(G_i|f_1,...,f_M)$ is calculated, where the probabilities f_1, …, f_M are substituted with $\hat{f}_1 *^{(t)},...,\hat{f}_M *^{(t)}$ in the tth iteration, beginning with the initial values $\hat{f}_1 *^{(0)},...,\hat{f}_M *^{(0)}$. In the M-step, the haplotype frequencies, $\hat{f}_m *^{(t+1)}$, are derived as

$$\hat{f}_m *^{(t+1)} = \frac{1}{2N} \sum_{i=1}^{N} \frac{\sum_{\eta\sim\gamma} \eta_m \hat{f}_\eta^{(t)} *}{\sum_{\eta\sim\gamma} f_\eta^{(t)} *} \text{ , for } m=1, \ldots, M.$$

The algorithm continues until a previously determined convergence criterion is fulfilled and the resulting approximate maximum likelihood estimator is denoted as $\hat{f}_1 *,...,\hat{f}_M *$. [Czika et al., 2002].

Based on the estimated haplotype frequencies, each subject is assigned the expected number of copies of a haplotype for m=1, …,M as:

$$H_{im} * = E[H_{im} \mid G_i = \gamma] = \frac{P\left(H_{im} = \eta_m, G_i = \gamma \middle| \hat{f}_1 *,...,\hat{f}_M *\right)}{P\left(G_i = \gamma \middle| \hat{f}_1 *,...,\hat{f}_M *\right)} = \frac{\sum_{\eta\sim\gamma} \eta_m \hat{f}_\eta *}{\sum_{\eta\sim\gamma} f_\eta *} \qquad (*).$$

3.1.3. Haplotype error measures

The accuracy of haplotype reconstruction can be measured in different ways for different purposes. A classification based on three characteristics is proposed: (1) The uncertainty across all haplotypes (1a, *overall error measure*) versus the error in a specific haplotype (1b, *haplotype-specific error measure*). (2) The uncertainty in a sample statistics (2a, i.e.: haplotype frequencies, $f \rightarrow f^*$) versus the uncertainty in individuals' haplotypes (2b). (3) To

further differentiate 2b: The error made by using the expected number of haplotype copies, $H \rightarrow H^*$ (3a), versus the error made by using the most probable haplotype, $H \rightarrow C^*$ (3b). The measures are summarized in Table 9 and defined and related to the above stated classes in the following.

Table 9: *Classification of measures for the haplotype reconstruction error.*

	Error in haplotype frequency	*Error in subject-specific haplotype* H_i[§]
Overall measure	Discrepancy (D)	- Error rate among all subjects (ER_{all}) - Error rate among ambiguous subjects (ER_{amb})
Haplotype -specific measure	Discrepancy per haplotype (D_m)	- Correlation between true and reconstructed haplotypes (R_m^2) - Sensitivity (Sn_m) and Specificity (Sp_m) - Misclassification probabilities

[§] H_i denotes the vector of length M coding the number of copies of true haplotypes of subject i, i=1, …, N for the m=1, …, M possible haplotypes.

3.1.3.1. Discrepancy

The *discrepancy D* is the average of the differences between true and reconstructed haplotype frequencies, providing an overall measure of the error $f \rightarrow f^*$ based on the summary statistics f instead of the subjects' haplotypes (class 1a, 2a):

$$D = D(f_1,...,f_M, f_1^*,...,f_M^*) = \frac{1}{2}\sum_{m=1}^{M}\left|f_m - f_m^*\right|.$$

The discrepancy is close to the mean squared error MSE [Adkins, 2004;Fallin and Schork, 2000], which is another way of averaging. A haplotype-specific discrepancy is given by $D_m(f_m, f_m^*) = \frac{1}{2}\left|f_m - f_m^*\right|$ for m=1,…, M (class 1b, 2a).

3.1.3.2. Error rate

The *error rate* among all individuals,

$$ER_{all} = \sum_{i=1}^{n}(1-c_i)/N, \qquad \text{where } c_i = \begin{cases} 0, H_i \neq C_i^* \\ 1, H_i = C_i^* \end{cases},$$

is the proportion of subjects with falsely classified haplotypes. Another definition is the error rate restricted to the subjects with ambiguous reconstruction, ER_{amb}, where N_{amb} replaces the

N in the denominator [Stephens et al., 2001]. ER_{all} and ER_{amb} are overall measures of the error $H \rightarrow C^*$ (class 1a, 2b, 3b).

3.1.3.3. Proportion of explained variance R^2_m

R_m^2, defined as the squared correlation between H_{im} and H_{im}^* [Stram et al., 2003b], $m=1, ..., M$, is a haplotype-specific measure for the error $H \rightarrow H^*$ *(class 1b, 2b, 3a)*. It is computed as the ratio of the haplotype variance explained by the genotypes, $\mathrm{Var}(H_{im}^*)$, to the variance of the $Bin(2, f_m)$-distributed (true haplotype frequency), $2\hat{f}_m^*\left(1-\hat{f}_m^*\right)$, assuming no error in the haplotype frequency from reconstruction ($f_m = f_m^*$).

3.1.3.4. Sensitivity and specificity

In the context of haplotypes, sensitivity and specificity are defined as "the probability that a true carrier of a certain haplotype is classified as such" (*sensitivity*) and "the probability that a true non-carrier is classified as such" (*specificity*), respectively, for $m=1, .., M$, $Sn_m = P(C_{im}^* > 0 \mid H_{im} > 0)$ and $Sp_m = P(C_{im}^* = 0 \mid H_{im} = 0)$. Thus, $1 - Sn_m$ and $1 - Sp_m$ measure the "haplotype-specific error" $H \rightarrow C^*$ (1b,2b,3b).

3.1.3.5. Misclassification probabilities

The error resulting from the transition $H \rightarrow C^*$ is a pure misclassification problem for a trichotomous variable, which is described by a 3x3 misclassification matrix consisting of the misclassification probabilities $\pi_{kl} = \Pr ob(C_{im}^* = k \mid H_{im} = l)$, $k,l = 0,1,2$ (Table 10a). Assuming no genotyping error, the subjects truly having two copies of a haplotype (true homozygous) as well as subjects with two copies of a haplotype in the reconstruction (observed homozygous) have always homozygous genotypes for all loci. These haplotypes can be reconstructed unambiguously and the misclassification probabilities π_{20}, π_{21}, π_{02}, and π_{12} equal zero. The misclassification matrix is then completely determined by sensitivity, specificity and the true haplotype probabilities or the observed haplotype probabilities, which is illustrated in Table 10b and c.

Table 10: *The misclassification matrix (a) in general form with* $\pi_{kl}^{(m)} = \Pr ob(C_{im}* = K \mid H_{im} = l)$, *k,l=0,1,2, m=1,…, M, for subject i, being the misclassification probabilities, (b) for the misclassification specific in haplotype reconstruction, expressed via sensitivity* $Sn_m = P(C_{im}* > 0 \mid H_{im} > 0)$, *specificity* $Sp_m = P(C_{im}* = 0 \mid H_{im} = 0)$ *and true haplotypes' probabilities (i.e. probabilities that a subjects has k number of copies of haplotype* h_m), $\pi_k^{(m)} = \Pr ob(H_{im} = k)$, *k=0,1,2,and (c) as in (b) but for the reconstructed haplotypes' probabilities* $\pi_k^{(m)}* = \Pr ob(C_{im}* = k)$. H_{im} *and* $C_{im}*$ *denote true and reconstructed number of copies of* h_m, *respectively.*

(a)

		Reconstructed $C_{im}*$			
		0	1	2	
True H_{im}	0	$\pi_{00}^{(m)}$	$\pi_{10}^{(m)}$	$\pi_{20}^{(m)}$	1
	1	$\pi_{01}^{(m)}$	$\pi_{11}^{(m)}$	$\pi_{21}^{(m)}$	1
	2	$\pi_{02}^{(m)}$	$\pi_{12}^{(m)}$	$\pi_{22}^{(m)}$	1

(b)

		Reconstructed $C_{im}*$			
		0	1	2	
True H_{im}	0	Sp	1-Sp	0	1
	1	$\pi_{01} = \frac{\pi_1^{(m)} + \pi_2^{(m)} - Sn(\pi_1^{(m)} + \pi_2^{(m)})}{\pi_1^{(m)}}$	$\pi_{11} = \frac{Sn(\pi_1^{(m)} + \pi_2^{(m)}) - \pi_2^{(m)}}{\pi_1^{(m)}}$	0	1
	2	0	0	1	1

(c)

		Reconstructed $C_{im}*$			
		0	1	2	
True H_{im}	0	Sp	1-Sp	0	1
	1	$\pi_{01} = \frac{-\pi_0* - Sp(\pi_0* - \pi_1* - \pi_2*) - Sn\pi_0*}{\pi_2* - \pi_0* + Sp(\pi_1* + \pi_0*) - Sn\pi_2*}$	$\pi_{11} = \frac{\pi_2* - Sp\pi_2* - Sn(\pi_0* + \pi_2*)}{\pi_2* - \pi_0* + Sp(\pi_1* + \pi_0*) - Sn\pi_2*}$	0	1
	2	0	0	1	1

3.1.4. Genotype frequencies from observed data

SNP data on numerous genes in a subsample of the population-based KORA study were available as examples. This sample of 704 individuals aged 55 to 74 years was a subset of the fourth survey (S4) of the KORA (Cooperative Research in the Region of Augsburg) study from 1999-2001 [Wichmann et al., 2005]. Genotypes were obtained via mass spectrometry (MALDI-TOF MS). The 8 genes in this investigation had been discussed as possible risk factors for diabetes, but had shown no or only a small association [Illig et al., 2003;Illig et al., 2004] : *IL-18*, *IL-13*, *MIP1A* , *INS*, *IL-6*, *MCP1*, *TNFA*, and *CAPN10*. Haplotypes were constructed and haplotypes frequencies derived by the EM algorithm (SAS proc haplotype) and also by PHASE. Depending on the gene, 2-7 loci were involved.

3.1.5. Simulation approach to quantify haplotype reconstruction error

In the simulations, true haplotype frequencies were taken as input parameters. For each simulation run, 1000 haplotypes were randomly drawn given the haplotype frequency distribution thus creating two copies of the haplotypes for 500 subjects assuming Hardy-Weinberg equilibrium (true haplotypes). Genotypes were deduced and haplotypes were reconstructed from these genotypes using the EM as well as the PHASE algorithm (reconstructed haplotypes). The reconstructed haplotypes were compared with the true haplotypes using the various error measures: For 100 simulations, the mean and the standard deviation of the error measures were computed. These mean error measures from the simulations were compared with analytical computations, which were also derived (see section 3.1.6).

To derive the true haplotype frequencies as input parameters, different scenarios were implemented for the simulations and the analytical computations:

<u>*Abstract scenarios*</u> including three types:

- *(a)* A two-locus scenario varying the frequency f_1 of haplotype h_1=00, while two other frequencies f_3 and f_4 are set at 0.1 and 0.05.
- *(b)* Another two-locus scenario varying the MAFs of locus 1 and locus 2, MAF_1 and MAF_2, and the correlation r. With $D_{LD} = r\sqrt{MAF_1 \cdot (1-MAF_1) \cdot MAF_2 \cdot (1-MAF_2)}$ [Devlin and Risch, 1995], the haplotype frequencies were derived as $f_1 = D_{LD} - MAF_1 \cdot MAF_2$,

$f_2 = -D_{LD} + (1 - MAF_1) \cdot MAF_2$, $f_3 = -D_{LD} + MAF_1 \cdot (1 - MAF_2)$, and $f_4 = D_{LD} - (1 - MAF_1) \cdot (1 - MAF_2)$.

(c) Various multi-locus scenarios assuming equal MAFs for 3-6 loci under the assumption of r=0.

Real data scenarios using the sets of haplotype frequencies as they were observed in the KORA data described above (section 3.1.4).

3.1.6. Analytical approach to quantify haplotype reconstruction error

Several haplotype error measures were also computed analytically.

3.1.6.1. Error rate

The probability that the haplotype pair of individual *i* is correctly inferred, $P(H_i = C_i^*)$, can be written as

$$\sum_{\eta} P(C_i^* = \eta, H_i = \eta) = \sum_{\eta} P(H_i = \eta) P(C_i^* = \eta \mid H_i = \eta) \quad (1)$$

with $P(C_i^* = \eta \mid H_i = \eta)$ rewritten as $P(C_i^* = \eta \mid H_i = \eta, G_i = \gamma) P(G_i = y \mid H_i = \eta)$. Since C_i^* depends only on G_i, the first factor reduces to $P(C_i^* = \eta \mid G_i = \gamma)$ and the second factor equals unity since G_i is uniquely defined by the haplotype pair H_i. Thus, with $\sum_{\eta:\eta\sim\gamma}$ being the sum over all possible haplotype pairs η, which are consistent with γ, $\eta \sim \gamma$, (1) can be restated as

$$\sum_{\gamma} \sum_{\eta:\eta\sim\gamma} P(H_i = \eta) P(C_i^* = \eta \mid G_i = \gamma) \quad (2).$$

The η among all $\eta \sim \gamma$ that maximize $P(H_i = \eta \mid G_i = \gamma)$ will be denoted by $\eta_{best}(\gamma) := \arg\max_{\eta:\eta\sim\gamma} P(H_i = \eta \mid G_i = \gamma)$, which is assigned as the most likely haplotype pair to C_i^* given the genotype. Thus, $P(C_i^* = \eta_{best}(\gamma) \mid G_i = \gamma) = 1$ and $P(C_i^* = \eta \mid G_i = \gamma) = 0$ for other $\eta : \eta \sim \gamma$, and (2) reduces to

$$\sum_{\gamma} P(H_i = \eta_{best}(\gamma)) \quad (3).$$

or written alternatively as $\sum_{\gamma} \max_{\eta:\eta\sim\gamma} P(H_i = \eta \mid G_i = \gamma)$. With f_η denoting the probability of a subject having haplotype pair η, this becomes $\sum_{\gamma} \max_{\eta:\eta\sim\gamma} f_\eta$. f_η is given by $f_\eta = 2 f_j f_k$ for $j \neq k$ or $f_\eta = (f_j)^2$ otherwise, when η depicts the haplotype pair h_j/h_k. If the reconstructed haplotype frequencies, $\hat{f}_\eta$ *, can be assumed to approximate the true haplotype frequencies (i.e. the haplotype-specific discrepancy is small), an approximation of the error rate is thus given by $1 - \sum_{\gamma} \max_{\eta:\eta\sim\gamma} \hat{f}_\eta$ *.

3.1.6.2. Sensitivity and specificity

For each haplotype h_m, m=1,…,M, the sensitivity and the specificity are defined as $Sn_m = P(C_{im}{}^* > 0 \mid H_{im} > 0)$ and $Sp_m = P(C_{im}{}^* = 0 \mid H_{im} = 0)$, respectively. The sensitivity can be rewritten as

$$Sn_m = \frac{P(C_{im}{}^* > 0, H_{im} > 0)}{P(H_{im} > 0)} \quad (5).$$

The numerator of (5) can be computed as $\sum_{\eta,\xi:\eta_m>0,\xi_m>0} P(C_i{}^* = \eta, H_i = \xi)$, which can be separated into two sums $\sum_{\eta=\xi:\eta_m>0,\xi_m>0} P(C_i{}^* = \eta, H_i = \xi) + \sum_{\eta\neq\xi:\eta_m>0,\xi_m>0} P(C_i{}^* = \eta, H_i = \xi)$, and due to $\sum_{\eta\neq\xi:\eta_m>0,\xi_m>0} P(C_i{}^* = \eta, H_i = \xi)$ being zero, one gets

$$\sum_{\eta:\eta_m>0} P(C_i{}^* = \eta, H_i = \eta) \quad (6).$$

With Bayes' Formula, (6) can be restated as $\sum_{\eta:\eta_m>0} P(H_i = \eta) P(C_i{}^* = \eta \mid H_i = \eta)$. Applying the same deduction as from (1) to (3), but restricting the sum $\sum_{\gamma}$ to a sum of γ, which yield $\eta_{best}(\gamma)$ such that the mth component of $\eta_{best}(\gamma)$ is > 0, (6) can be restated as

$$\sum_{\gamma\eta_{best}(\gamma)_m>0} P(H_i = \eta_{best}(\gamma)) \quad (7),$$

and the sensitivity can be derived as

$$Sn_m = \frac{\sum_{\gamma:\eta_{best}(\gamma)_m>0} P(H_i = \eta_{best}(\gamma))}{\sum_{\eta:\eta_m>0} P(H_i = \eta)} = \frac{\sum_{\gamma:\eta_{best}(\gamma)_m>0} f_{\eta_{best}}}{\sum_{\eta:\eta_m>0} f_\eta}$$

and the specificity, analogously, as

$$Sp_m \frac{\sum_{\eta_{best}(\gamma)_m=0} P(H_i = \eta_{best}(\gamma))}{\sum_{\eta:\eta_m=0} P(H_i = \eta)} = \frac{\sum_{\eta:\eta_{best}(\gamma)_m=0} f_{\eta_{best}}}{\sum_{\eta:\eta_m=0} f_\eta}.$$

These terms were implemented in the R program in the function "Sensitivity" (see Appendix A.1.).

3.1.6.3. Computing the error rate from sensitivity and specificity

In order to describe the error rate $1 - P(H_i = C_i^*) = 1 - \sum_\eta P(C_i^* = \eta, H_i = \eta)$ by Sn_m and Sp_m, the sum $\sum_\eta$ is partitioned into $\sum_{\eta:\eta_m>0}$ and $\sum_{\eta:\eta_m=0}$. With (5) and (6), one obtains $\sum_{\eta:\eta_m>0} P(C_i^* = \eta, H_i = \eta) = Sn_m \sum_{\eta:\eta_m>0} P(H_i = \eta)$ and $\sum_{\eta:\eta_m=0} P(C_i^* = \eta, H_i = \eta) = Sp_m \sum_{\eta:\eta_m=0} P(H_i = \eta)$, and thus $P(H_i = C_i^*) = Sp_m \sum_{\eta:\eta_m=0} P(H_i = \eta) + Sn_m \sum_{\eta:\eta_m>0} P(H_i = \eta) = Sp_m \sum_{\eta:\eta_m=0} f_\eta + Sn_m \sum_{\eta:\eta_m>0} f_\eta$.

3.1.6.4. Example for two loci: Analytical computation of sensitivity and specificity

The four possible haplotypes are h_1=00, h_2=01, h_3=10 and h_4=11, with 0 or 1 denoting the major or the minor allele, respectively. A subject i with the genotypes 0/0 and 1/1 at the two loci is coded as G_i=(0,2); such a subject has definitely two copies of the haplotype h_2, H_i is then (0,2,0,0), and H_i* and C_i* equal H_i (no reconstruction error). A subject with the genotypes 0/1 and 0/1 is coded as G_i =(1,1) and has either the haplotype pair h_1/h_4 or h_2/h_3. H_i is thus (1,0,0,1) or (0,1,1,0). For this genotype and a given set of $\hat{f}_1^*, \hat{f}_2^*, \hat{f}_3^*, \hat{f}_4^*$, e.g. the first component of H_i*, H_{i1}*, is computed according to (*) in section 3.1.2 as

$$\frac{1 \cdot 2 \cdot \hat{f}_1^* \cdot \hat{f}_4^* + 0 \cdot 2 \cdot \hat{f}_2^* \cdot \hat{f}_3^*}{2 \cdot \hat{f}_1^* \cdot \hat{f}_4^* + 2 \cdot \hat{f}_2^* \cdot \hat{f}_3^*} = \frac{\hat{f}_1^* \cdot \hat{f}_4^*}{\hat{f}_1^* \cdot \hat{f}_4^* + \hat{f}_2^* \cdot \hat{f}_3^*} = \frac{1}{1 + \frac{\hat{f}_2^* \cdot \hat{f}_3^*}{\hat{f}_1^* \cdot \hat{f}_4^*}}$$

For the haplotype frequencies $(\hat{f}_1^*, \hat{f}_2^*, \hat{f}_3^*, \hat{f}_4^*) = (0.3, 0.3, 0.25, 0.15)$ H_i* turns out to be (0.375,0.625,0.625,0.375) with the categorization yielding C_i* =(0,1,1,0) (thus no reconstruction error when viewing $H_i \rightarrow C_i$*). In another sample with other sample haplotype frequencies $(\hat{f}_1^*, \hat{f}_2^*, \hat{f}_3^*, \hat{f}_4^*) = (0.4, 0.2, 0.25, 0.15)$ such a subject is assigned

H_i*=(0.545,0.455,0.455,0.545) and thus C_i* =(1,0,0,1) (hence a misclassification when viewing $H_i \rightarrow C_i$*). Thus, the haplotype C*=(1,0,0,1) would be inferred for all subjects with this genotype in the sample, if $\hat{f}_1 * \cdot \hat{f}_4 * > \hat{f}_2 * \cdot \hat{f}_3 *$, or C*=(0,1,1,0) otherwise. In the two locus case, the ambiguity fraction can be calculated as P(h_1/h_4) + P(h_2/h_3) = $2 \cdot \hat{f}_1 * \cdot \hat{f}_4 * + 2 \cdot \hat{f}_2 * \cdot \hat{f}_3 *$, yielding an ambiguity fraction of 0.24 for the first example of haplotype frequencies and 0.22 for the second.

Table 11 can be used to calculate the sensitivity and specificity, e.g. for haplotype h_1,

$$Sn_1 = \frac{f_1 \cdot f_1 + 2 \cdot f_1 \cdot f_2 + 2 \cdot f_1 \cdot f_3 + I_{(f_1 \cdot f_4 > f_2 \cdot f_3)} \cdot (2 \cdot f_1 \cdot f_4)}{f_1 \cdot f_1 + 2 \cdot f_1 \cdot f_2 + 2 \cdot f_1 \cdot f_3 + 2 \cdot f_1 \cdot f_4} \text{ and}$$

$$Sp_1 = \frac{f_2 \cdot f_2 + I_{(f_2 \cdot f_3 > f_1 \cdot f_4)} \cdot (2 \cdot f_2 \cdot f_3) + 2 \cdot f_2 \cdot f_4 + f_3 \cdot f_3 + 2 \cdot f_3 \cdot f_4 + f_4 \cdot f_4}{f_2 \cdot f_2 + 2 \cdot f_2 \cdot f_3 + 2 \cdot f_2 \cdot f_4 + f_3 \cdot f_3 + 2 \cdot f_3 \cdot f_4 + f_4 \cdot f_4}$$ with I being the indicator function yielding the value 1 when the condition holds.

Table 11: *Two-locus example: Haplotype pairs, their probabilities, the resulting genotypes, and the genotype probability for the two-locus case. The frequencies of the haplotypes h_1, h_2, h_3 and h_4 are denoted with f_1, f_2, f_3, and f_4. For the ambiguous genotype G=(1,1) (shaded in gray), the haplotype pair with the higher probability (i.e. (1,0,0,1) if $f_1f_4 > f_2f_3$, or (0,1,1,0) else) is chosen for all subjects in the sample.*

	h_1=00	h_2=01	h_3=10	h_4=11
h_1=	P(H=(2,0,0,0))=f_1f_1	P(H=(1,1,0,0))=$2f_1f_2$	P(H=(1,0,1,0))=$2f_1f_3$	P(H=(1,0,0,1))=$2f_1f_4$
00	P(G=(0,0))=f_1f_1	P(G=(0,1))=$2f_1f_2$	P(G=(1,0))=$2f_1f_3$	P(G=(1,1))=$2(f_1f_4+f_2f_3)$
h_2=		P(H=(0,2,0,0))=f_2f_2	P(H=(0,1,1,0))=$2f_2f_3$	P(H=(0,1,0,1))=$2f_2f_4$
01		P(G=(0,2))=f_2f_2	P(G=(1,1))=$2(f_1f_4+f_2f_3)$	P(G=(1,2))=$2f_2f_4$
h_3=			P(H=(0,0,2,0))=f_3f_3	P(H=(0,0,1,1))=$2f_3f_4$
10			P(G=(2,0))=f_3f_3	P(G=(2,1))=$2f_3f_4$
h_4=				P(H=(0,0,0,2))=f_4f_4
11				P(G=(2,2))=f_4f_4

3.2. Results

3.2.1. Discrepancy

Figure 13a illustrates that the discrepancy increases steadily with increasing frequency of one haplotype (other frequencies fixed, *abstract scenarios type a*) until it reaches a maximum of 0.00917 for f_1=0.46, and then, for f_1>0.61, it decreases monotonically. The discrepancies in *real data scenarios* (Table 12) show values below 0.005 indicating an average difference between true and reconstructed haplotype frequencies below 0.5%, except for *MCP1* and *CAPN10*. There is a small difference when comparing EM or PHASE derived haplotypes yielding a smaller discrepancy using the EM for *INS* and *MCP1*, and a smaller discrepancy using PHASE for *IL-6*.

Table 12: *Real data scenarios: Discrepancy D ± standard deviation using the EM- or PHASE-reconstruction, stating the number of effective loci, l_{eff}, and the number of loci, l.*

l_{eff}	l	*Gene*	*EM*	*PHASE*
1.02	2	*IL-18*	0.0000 ± 0.0000	0.0000 ± 0.0000
1.06	3	*IL-13*	0.0001 ± 0.0003	0.0001 ± 0.0003
1.15	2	*MIP1A*	0.0002 ± 0.0004	0.0002 ± 0.0004
1.69	4	*INS*	0.0003 ± 0.0007	0.0005 ± 0.0006
2.31	3	*IL-6*	0.0008 ± 0.0011	0.0006 ± 0.0007
2.96	3	*MCP1*	0.0130 ± 0.0070	0.0150 ± 0.0080
3.00	3	*TNFA*	0.0040 ± 0.0030	0.0040 ± 0.0030
6.38	7	*CAPN10*	0.0320 ± 0.0090	0.0320 ± 0.0100

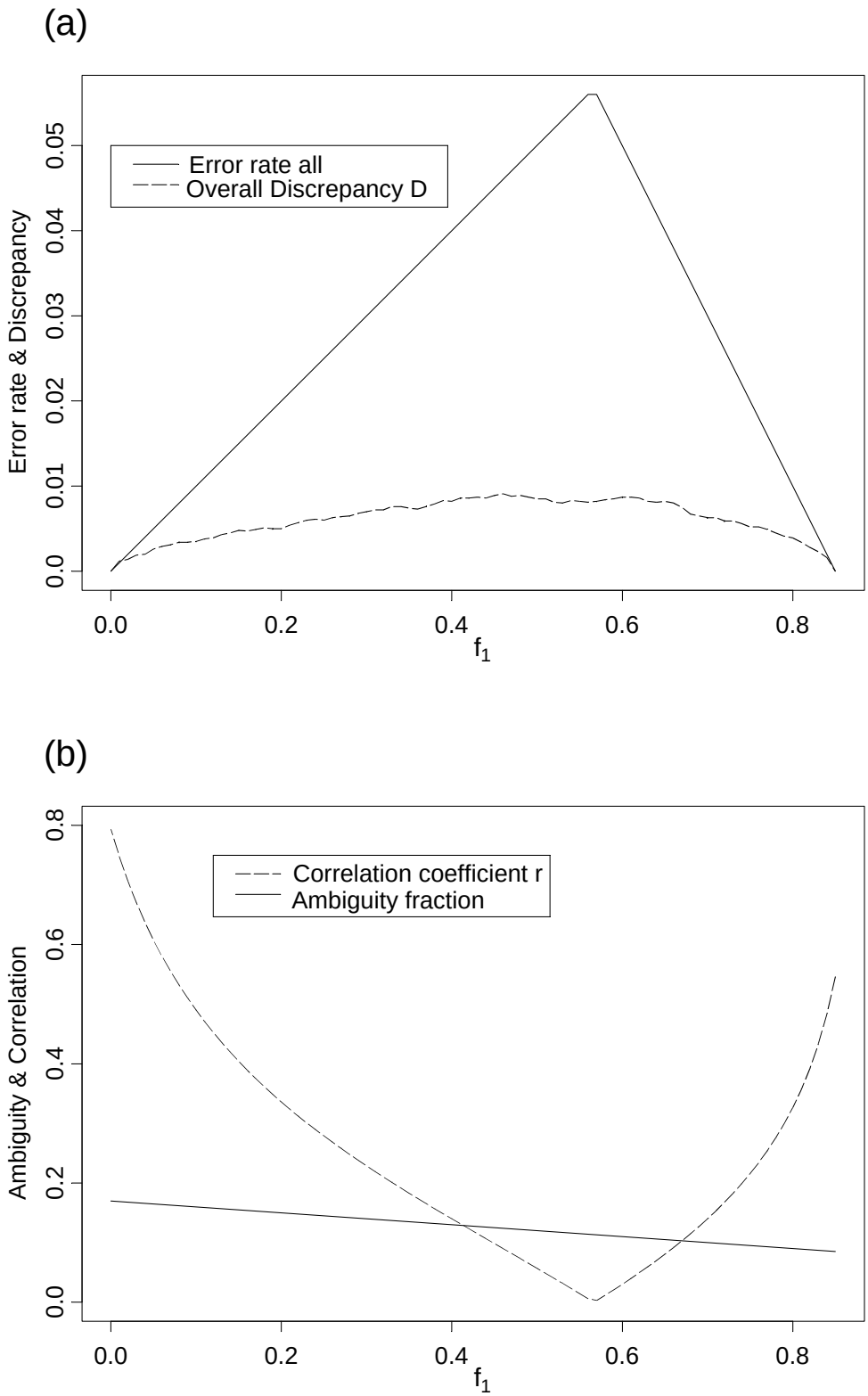


Figure 13: *Discrepancy and error rate depending on haplotype frequency: (a) Discrepancy (from simulations) and error rate (analytically derived), (b) ambiguity fraction and correlation coefficient r (Abstract type a scenarios: two loci varying frequency f_1 of haplotype h_1=00 with f_3 =0.1 and f_4 =0.05 for h_3=01 and h_4=11).*

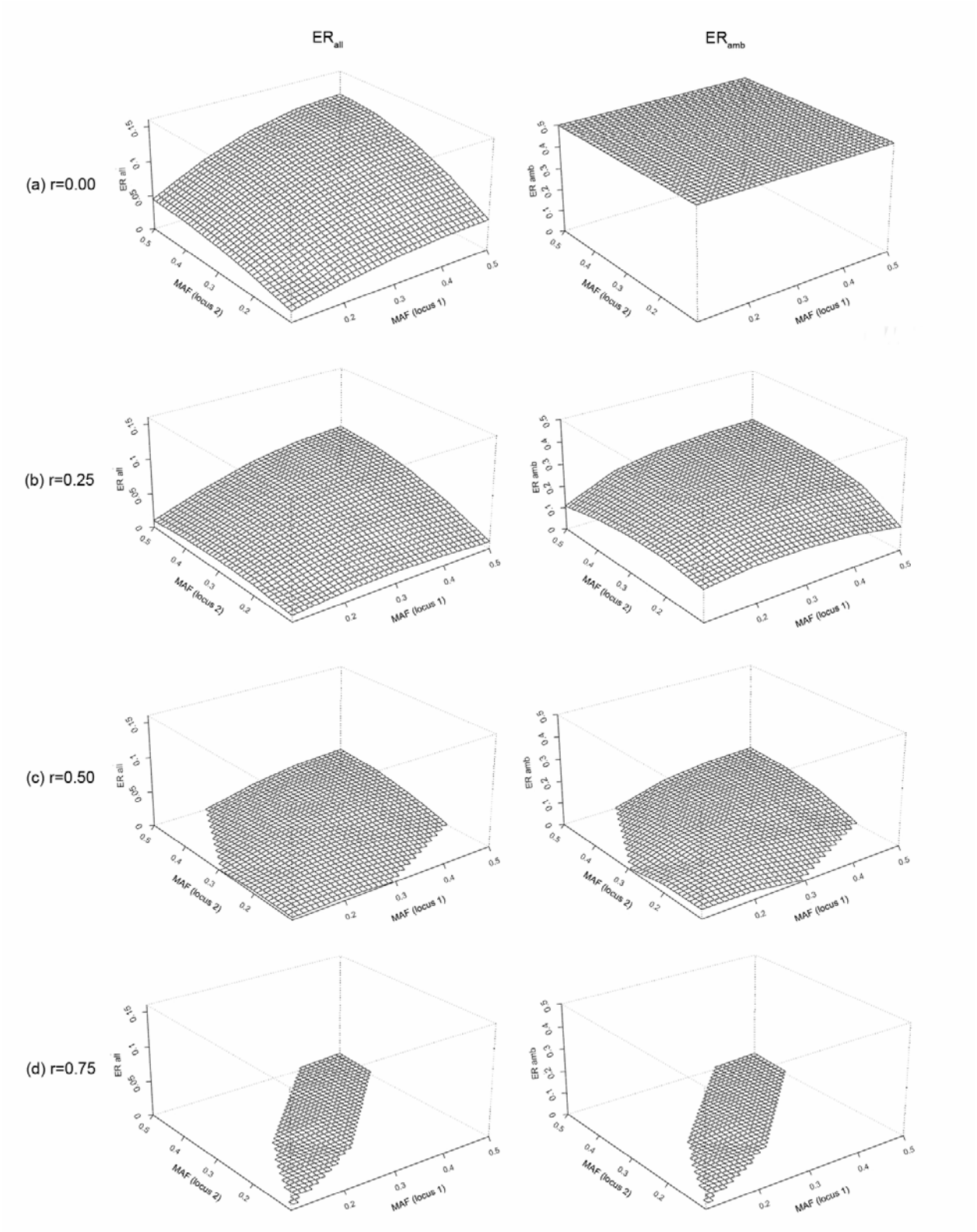


Figure 14: *Error rate for varying MAF and correlations for two loci: Analytically derived error rate for (a) r = 0, (b) r = 0.25, (c) r = 0.5, (d) r = 0.75 (Abstract type b scenarios)*

3.2.2. Error rate

For the two-locus scenario varying one haplotype frequency (*abstract scenarios type a*), the error rate (see Figure 13a) reaches a maximum of 0.056 for f_1=0.57 and is minimal for small f_1 or for large f_1. To understand the location of this maximum, the ambiguity and the correlation coefficient *r* are displayed in Figure 13(b): At the maximum, the alleles show zero correlation. It also becomes apparent that the correlation has a stronger influence on the error rate than the ambiguity in this 2-locus case.

Figure 14 depicts the dependency of the error rate on the MAFs and r (*abstract scenario type b*): the smaller the MAF, the smaller the error. The error is minimal, when both MAFs are small. This is due to the fact that the ambiguity is smaller for lower MAF as there are fewer genotypes deviating from the wildtype and thus fewer heterozygotes. Furthermore, the error decreases for increasing *r*. Note that high correlation imposes strong restrictions on the possible MAF combinations, as the MAF of the second locus can only slightly deviate from the MAF of the first locus, and thus the parameter space is reduced (Figure 14d). When *r*=0, ER_{amb} is 0.5, which is like flipping a coin (Figure 14a) for assigning haplotypes to ambiguous subjects.

(a) (b) (c)

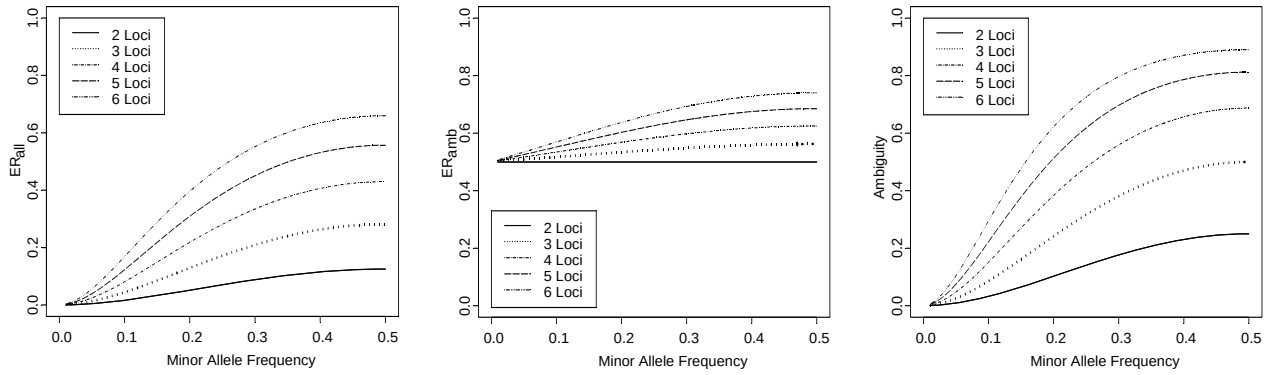


Figure 15: *Error rate for varying number of loci and MAF under no correlation: (a) ER_{all} and (b) ER_{amb} and (c) ambiguity analytically derived for abstract type c scenarios (2-6 loci, r=0, equal MAF at each locus).*

In Figure 15a, depicting the multi-locus scenarios under no LD (*abstract scenarios type c*), it can be seen that the error rate increases with the number of loci. This is due to the fact that the probability of a subject being heterozygous in at least two loci increases with the number of loci involved, which is depicted by the increasing ambiguity fraction (Figure 15c). But this is

not the sole reason as it can be seen that also the ER_{amb} increases with the number of loci (Figure 15b): The number of haplotypes also increases with the number of loci and thus the pool for misclassification enlarges. It can be further seen, that, for MAF=0.5 when all alleles and consequently all haplotypes are equally frequent, the ER_{amb} is as large as when a die was rolled for haplotype assignment of ambiguous subjects. This is due to the fact that then the haplotype inference is guided by neither the correlation nor the haplotype frequency. When MAF<0.5 and haplotypes occur with different frequencies, the reconstruction can improve by preferring haplotype pairs containing more frequent haplotypes. Note that Figure 15 shows a worst-case scenario indicating the maximum possible error due to the no-LD assumption. As it is unreasonable to infer haplotypes in such a situation in the first place, these error rates remain unmatched in real data scenarios.

Table 13 shows that the error rates for real data scenarios varies substantially between genes.

Table 13: *Real data scenarios: Error rate (ER_{all}) and the error rate among ambiguous subjects (ER_{amb}) ± standard deviation derived from simulations with EM-reconstruction, PHASE-reconstruction, as well as the error rate computed by the analytical approach given in the Appendix.*

l_{eff}	l	gene	*Simulations using EM*		*Simulations using PHASE*		*Analytical Approach*	
			ER_{amb}	ER_{all}	ER_{amb}	ER_{all}	ER_{amb}	ER_{all}
1.023	2	*IL-18*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.064	3	*IL-13*	0.0001 ± 0.0008	0.0000 ± 0.0003	0.0001 ± 0.0008	0.0000 ± 0.0003	0.0001	0.0000
1.149	2	*MIP1A*	0.0002 ± 0.001	0.0001 ± 0.0004	0.0000 ± 0.0010	0.0001 ± 0.0004	0.0002	0.0001
1.687	4	*INS*	0.0007 ± 0.0019	0.0003 ± 0.0008	0.0006 ± 0.0017	0.0003 ± 0.0007	0.0002	0.0001
2.313	3	*IL-6*	0.0020 ± 0.003	0.0008 ± 0.0015	0.0005 ± 0.0014	0.0003 ± 0.0007	0.0003	0.0001
2.959	3	*MCP1*	0.2560 ± 0.046	0.048 ± 0.01	0.2580 ± 0.051	0.48 ± 0.10	0.2460	0.0460
2.999	3	*TNFA*	0.456 ± 0.166	0.01 ± 0.0040	0.428 ± 0.1670	0.10 ± 0.004	0.3900	0.0090
6.384	7	*CAPN10*	0.199 ± 0.024	0.125 ± 0.015	0.197 ± 0.0240	0.123 ± 0.015	0.1870	0.1170

The error rate is large for genes with low LD between loci, which are the genes showing a small difference between the number of loci and the effective number of loci (e.g. for *MCP1*, *TNFα*, *CAPN10*). For most genes, the error rate is well below 1%, which indicates that for 99% of the subjects the haplotypes are perfectly reconstructed. In these real data scenarios, the error rate of PHASE-reconstructed haplotypes is very similar to EM-based haplotypes; the analytical approach yields similar results as the simulation approach, but slightly lower.

3.2.3. Haplotype specific error measures

A starplot was developed and implemented in the R function "starplot" (see Appendix A.2.) to summarize the haplotype-specific errors: Figure 16 and Figure 17 show star plots for three selected genes with < 5 loci and for the *CAPN10* gene involving 7 loci (*real data scenarios*). The measures were derived analytically (see section 3.1.6), but were very similar in the simulations. Comparing Figure 16 with Table 13 shows that high R^2 appears together with high sensitivity, and that low error rate occurs with high R^2 and high sensitivity.

Furthermore, it can be seen that the specificity is reduced rather for common haplotypes (e.g. 98% for haplotype 000 for *TNFα*, 97% for 0000000 of *CAPN10*). On the other side, the sensitivity is reduced rather for rare haplotypes (e.g. 101 of *MCP1)*. However, there are also rare haplotypes which show almost perfect sensitivity (e.g. 0010101 for *CAPN10*).

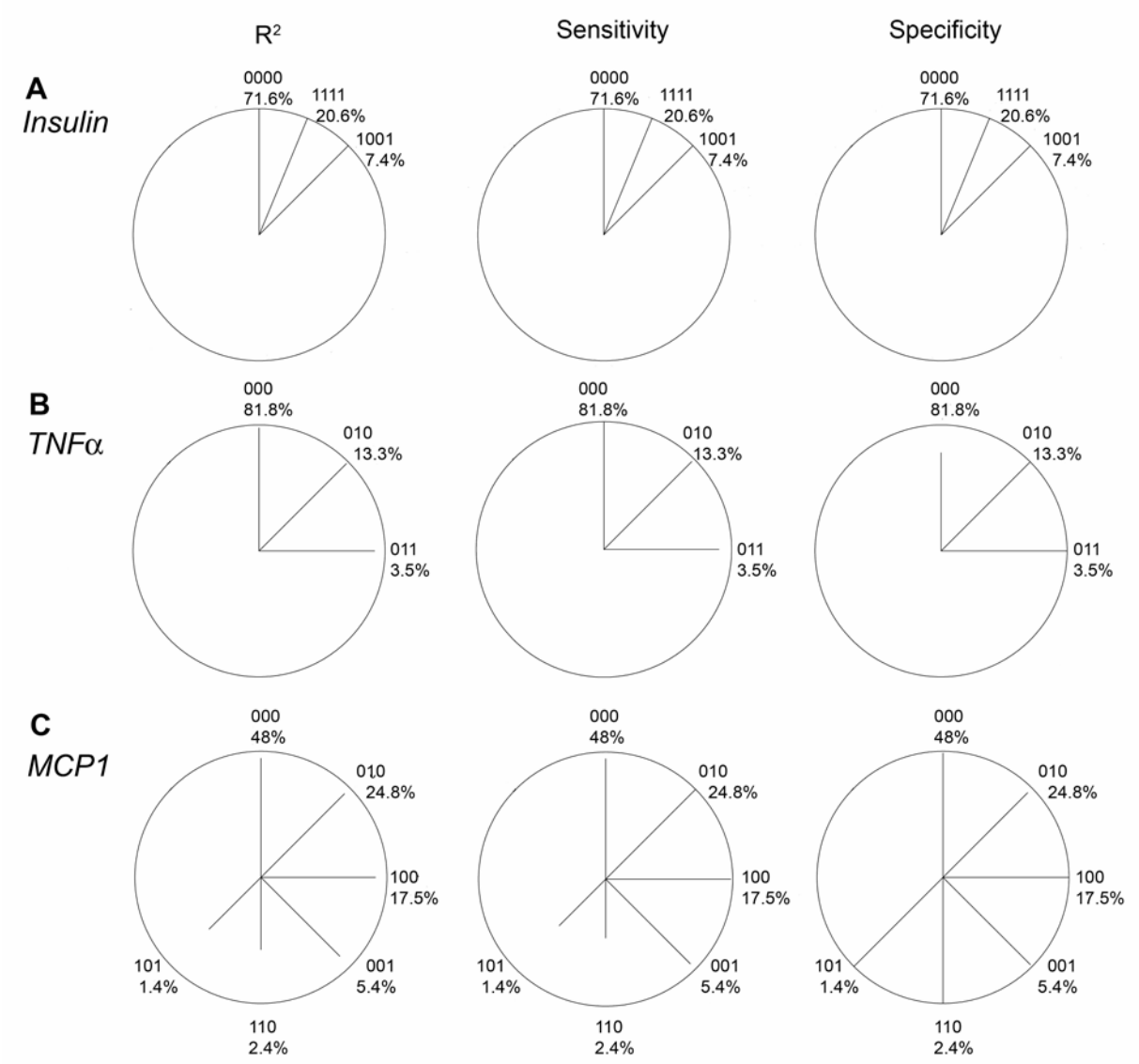


Figure 16: *Haplotype-specific error measures: Star plots for various genes displaying R^2, sensitivity and specificity (analytically derived) as the length of the line for each common haplotype (frequency ≥1%). A line reaching the circle indicates a value of 100% (no error). Haplotypes are labeled using 0/1 coding for major/minor allele and stating the haplotype frequency. Lines are sorted clockwise by haplotype frequency beginning at the top with the most frequent haplotype. The angle between lines is given by the number of possible haplotypes, i.e. $360°/2^L$, where L is the number of loci. The proportion without lines thus indicates the proportion of rare or non-existing haplotypes (<1% frequency).*

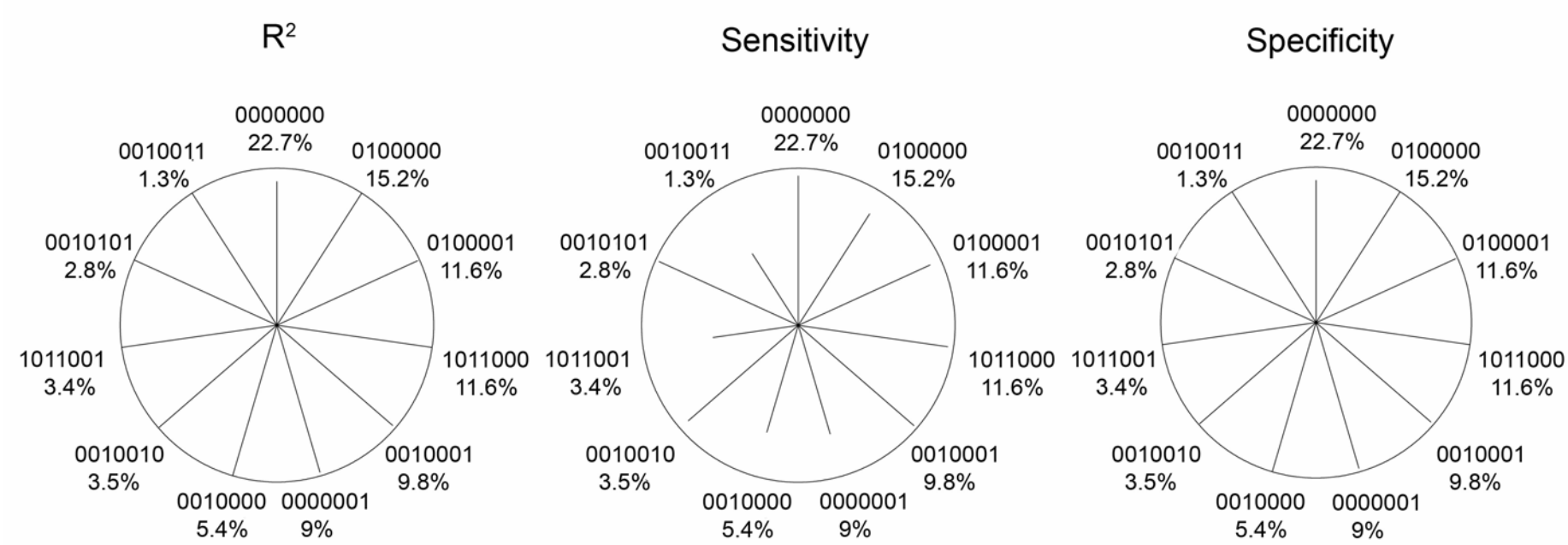


Figure 17: *Haplotype-specific error measures: Star plots for the CAPN10 displaying the R^2, the sensitivity, or the specificity as in Figure 16. The angle between lines is given by 360° divided by the number of frequent haplotypes (frequency ≥1%) to accommodate for the large number of loci (L=7).*

4. Impact of haplotype misclassification from genotype error and reconstruction on association analysis

After looking at misclassification in genotypes and haplotype reconstruction separately, both error sources are now combined to evaluate the overall misclassification in haplotypes, if genotype error can be assumed and haplotypes have to be reconstructed out of these error-prone genotypes. A resampling method was used to estimate haplotype misclassification matrices. Then, the impact of this haplotype misclassification on association estimates is evaluated in the example of the *APM1* gene using the MC-SIMEX approach.

In the following, the genotype error is not restricted to errors induced by the genotyping process, which can be estimated from repeated genotyping of the same samples as in section 2, but may result also from other sources such as diluted DNA.

4.1. Methods and material

4.1.1. Misclassification from genotype and haplotype error combined

As in chapter 2 and 3, true genotypes are denoted as $G_i = (G_{i1},...,G_{iL})$ for the ith subject, i=1,…, n, for L SNPs with G_{il} indicating the number of minor alleles at locus l for individual i ($G_{il} \in \{0,1,2\}$) and the genotype probabilities $\pi_G^{(l)} = (\pi_{G,0}^{(l)}, \pi_{G,1}^{(l)}, \pi_{G,2}^{(l)})$. The observed genotypes derived in the laboratory and potentially subject to genotype error, are denoted as $G^*_i = (G^*_{i1},...,G^*_{iL})$ at L loci with $\pi_G^{(l)}* = (\pi_{G,0}^{(l)}*, \pi_{G,1}^{(l)}*, \pi_{G,2}^{(l)}*)$, which can be estimated by observed genotype frequencies.

An allele-independent error model for the genotype misclassification is applied here, since it is the most parsimonious model without loosing too much information. The probability of misclassifying the major allele A as the minor allele a equals the probability of misclassifying the minor allele as the major, $P(A \to a) = P(a \to A) = \varepsilon$. Values of 0.005 and 0.01 are applied for ε. Note, that ε was chosen slightly higher than the parameter estimated in the second chapter of this work (ε = 0.000997, section 2.2.3). This estimated misclassification parameter can only approximate the error induced by the genotyping process itself but cannot account for errors from other sources. Therefore, ε-values of 0.005 and 0.01 should be in the ballpark of error rates expected if nowadays state-of-the-art genotyping methods are used.

As in 3.1, the transition $H_{im} \rightarrow C_{im}$ * from the true to the most likely haplotype forms a classical 3x3 misclassification problem which can be described by the misclassification matrix $\Pi^{(m)} = (\pi_{kl}^{(m)})_{k,l=0,1,2}$, that is the matrix of the misclassification probabilities $\pi_{kl}^{(m)} = P(C_{im}* = k \mid H_{im} = l)$, $k,l = 0,1,2$. Since $\pi_{0l}^{(m)} + \pi_{1l}^{(m)} + \pi_{2l}^{(m)} = 1$ for $l = 0,1,2$, the misclassification matrix involves 6 unknown parameters.

In the case of no genotype error, the subjects truly having two copies of a 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 $\pi_{21}^{(m)}$ equal zero. In this case, the misclassification problem reduces to two unknown parameters $\pi_{00}^{(m)}$ and $\pi_{01}^{(m)}$ with $\pi_{11}^{(m)} = 1 - \pi_{00}^{(m)} - \pi_{01}^{(m)}$. This can be reparameterized for the sensitivity and the specificity as in Table 10.

The misclassification problem *H*→*C** also involves the genotype error *G*→*G** (see Figure 8). Haplotypes, that would have been unambiguous through statistical reconstruction alone are now also subject to error due to the genotype error and $\pi_{02}^{(m)}$, $\pi_{12}^{(m)}$, $\pi_{20}^{(m)}$ and $\pi_{21}^{(m)}$ may deviate from zero leaving 6 parameters for misclassification estimation. The sensitivity and specificity can then be determined from the misclassification probabilities as $Sn_m = \dfrac{\pi_{11}^{(m)}\pi_1^{(m)} + \pi_{12}^{(m)}\pi_2^{(m)} + \pi_{21}^{(m)}\pi_1^{(m)} + \pi_{22}^{(m)}\pi_2^{(m)}}{\pi_1^{(m)} + \pi_2^{(m)}}$ and $Sp_m = \pi_{00}^{(m)}$.

If a dominant genetic effect model is assumed, the 3x3-misclassification matrix is reduced to a 2x2 problem and thus, again, completely determined by sensitivity and specificity, even if genotype error is involved.

4.1.2. Approximating the haplotype misclassification matrix via resampling

The estimated probabilities of the haplotypes, $f = (f_1,...,f_M)$, are assumed to be the true haplotype probabilities and set as input parameters. This approach is suitable, since haplotype specific discrepancies have been shown to be very small (section 3.2.1). For each simulation run, 1000 haplotypes were randomly drawn given the haplotype probability distribution. Two haplotypes are randomly assigned to each of 500 subjects assuming random

mating (Hardy-Weinberg equilibrium). From these true haplotypes of each subject, genotypes were deduced. The genotypes were then subject to genotype error with error size ε=0%, 0.5% and 1% for each allele, which implies that a subject with a true homozygous genotype at one SNP is assigned a heterozygous genotype with probability $2\varepsilon(1-\varepsilon)$ and a homozygous genotype of the other allele with probability ε^2. From these observed error-prone genotypes, haplotypes were then reconstructed using the EM algorithm as implemented in SAS proc haplotype. The reconstructed haplotypes were 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 the error measures were computed.
Note, that error measures derived by this approach can only approximate the expected error, since there are no repeated measurements as well as gold standard methods available to estimate haplotype misclassification from combined genotype error and haplotype reconstruction (for error estimation methods see 1.3.1).

4.1.3. Simulations to evaluate bias in haplotype association estimates and MC-SIMEX performance

To evaluate the applicability of a method, which corrects for errors on a specific situation, simulations have to be conducted. Only in this case, true as well as observed parameters are known and can be compared. To assess the performance of the MC-SIMEX in haplotype association analysis, normally distributed outcome data for linear regression analysis were simulated. A specific haplotype h_R was assumed to be the risk haplotype of interest and all other haplotypes were not distinguished and denoted here as h. The true probability of the risk haplotype was denoted by f_{h_R}.

For the dominant model, subjects were randomly assigned the wildtype haplotype-pair ("h/h") with probability $(1-f_{h_R})^2$ and an outcome variable from $N(0,\sigma^2)$. For subjects with at least one copy of the risk haplotype ("h_R/h" or "h_R/h_R") the outcome variable was randomly drawn from $N(\beta_{dom},\sigma^2)$. For the additive model, subjects were assigned the "h/h" haplotype with probability $(1-f_{h_R})^2$, the "h_R/h" with probability $2f_{h_R}(1-f_{h_R})$ and "h_R/h_R" with probability ${f_{h_R}}^2$ and the corresponding outcome variables were drawn from $N(0,\sigma^2)$, $N(\beta_{add},\sigma^2)$, or $N(2\beta_{add},\sigma^2)$, respectively. Thus, the measure of effect was the difference in the mean value

of the outcome for each additional copy of h_R compared to subjects with no copy of h_R (β_{add}) or the difference between subjects having at least one h_R compared to no h_R (β_{dom}). The variance of the outcome, σ^2, was assumed to be equal for all subjects and set to 0.4 to mimic the variance of adiponectin plasma level in the real data example. Different scenarios were applied: β_{add} and β_{dom} were set to 0.5 or 0.05; f_{h_R} of 0.15 and 0.3 was used. Based on the assigned true haplotypes for each subject, H_{im}, and the given misclassification probabilities, observed haplotypes, H_{im}, were derived. Two different misclassification schemes were used, depicting a rather low (Π_{low}) as well as a high misclassification (Π_{high}):

$$\Pi_{low} = \begin{pmatrix} 0.975 & 0.1 & 0.01 \\ 0.025 & 0.9 & 0.1 \\ 0 & 0 & 0.89 \end{pmatrix} \text{ and } \Pi_{high} = \begin{pmatrix} 0.899 & 0.3 & 0.1 \\ 0.1 & 0.69 & 0.3 \\ 0.001 & 0.01 & 0.6 \end{pmatrix}$$

The dimension of both misclassification matrices was based on ranges of misclassification matrices estimated in a real data example (Table 14).

The most probable number of haplotypes C_{im}* were coded as Dummy-variables and thus, for each haplotype, estimates are obtained for the heterozygous (C_{im}*=1: β_1) as well as homozygous rare haplotype pair (C_{im}*=2: β_2) comparing "h_R/h" with "h/h" or "h_R/h_R" with "h/h". Assuming the dominant model, only one parameter is estimated, so that $\hat{\beta}_{dom} = \beta_{>=1}$.

For each of the 200 performed simulations with 1000 subjects each, the effect estimate was computed ignoring the error in the haplotypes (naive estimate) as well as the SIMEX corrected effect estimates applying a linear, quadratic and log-linear function for extrapolation. Mean and standard deviation (for simulation precision) of estimates were derived as well as the 95% coverage, which describes the proportion of 95% confidence intervals that contain the true effect.

4.1.4. Correction of association from *APM1* haplotypes on adiponectin

As in the part on genotyping error (chapter 2), genotype data from *APM1* gene (section 1.2.6) are used for a realistic data example. Haplotypes were reconstructed based on the 15 selected haplotype tagging SNPs [Heid et al., 2006] via the EM algorithm (SAS proc haplotype). The corresponding estimated haplotype frequencies were then used to approximate different error measures and misclassification matrices for each haplotype as described above (4.1.2). Linear regression estimates were first computed ignoring the error in the haplotypes (naive estimates)

as well as MC-SIMEX corrected estimates with the log-linear extrapolation function using the approximated misclassification matrices. For haplotype association, the linear model was computed on log(adiponectin+1) and adjusted for age, sex, body-mass index (BMI) and all other haplotypes with the most frequent one as reference (H22, frequency=0.124). Since *APM1*-haplotypes are rather rare, subjects homozygous for one haplotype were coded in the same group as subjects having one copy of this very haplotype, which corresponds to a dominant model, yielding one coefficient β. Only for haplotypes with more than 10 homozygous subjects, additional estimates were obtained separately for one copy (β_1) as well as two copies (β_2) of this haplotype. Effects based on the expected values of haplotypes H* (β_{H*}) were also derived for these haplotypes, implying the assumption of additivity, which is known as haplotype trend regression [Zaykin et al., 2002]. The misclassification probabilities for this problem were derived by the simulations described above assuming a genotype error of 0% (pure reconstruction error), 0.5% and 1%.

4.2. Results

4.2.1. Quantification of the Haplotype misclassification problem

4.2.1.1. Estimated misclassification matrix for haplotype misclassification

In order to describe the error in the haplotypes from genotyping error and haplotype reconstruction combined, haplotype misclassification probabilities were derived for a specific SNP data set. Haplotype frequencies and the genotype error were given via the simulations exemplified on the *APM1* data. Table 14 (and Appendix A.3.) depicts the haplotype misclassification matrices without and with genotype error, for a selection of three *APM1* haplotypes. It can be seen that π_{02}, π_{12}, π_{20} and π_{21} are zero in the case of no genotype error, as already stated in the methods section. When adding a genotype error with ε= 0.5% or 1%, all misclassification probabilities (for $i \neq j$) increase and most of the probabilities π_{02}, π_{12}, π_{20} and π_{21} deviate from zero. Note, that the influence of the genotype error of 0.5% or 1% on the overall misclassification on top of the haplotype reconstruction error is partially larger than the pure reconstruction error itself (genotype error=0%).

***Table 14**: Misclassification matrices for selected APM1 haplotypes assuming 0, 0.5 and 1% genotype error per allele*

Genotype Error

0%

Haplotype H22, Frequency 0.124

		H		
		0	1	2
C*	0	0.9644	0.0275	0
	1	0.0356	0.9725	0
	2	0	0	1

Haplotype H2, Frequency 0.053

		H		
		0	1	2
C*	0	0.9947	0.0070	0
	1	0.0053	0.9930	0
	2	0	0	1

Haplotype H12, Frequency 0.023

		H		
		0	1	2
C*	0	0.9947	0.1030	0
	1	0.0053	0.8970	0
	2	0	0	1

0.5%

Haplotype H22, Frequency 0.124

		H		
		0	1	2
C*	0	0.9560	0.0890	0.0025
	1	0.0440	0.9105	0.1275
	2	0	0.0004	0.8700

Haplotype H2, Frequency 0.053

		H		
		0	1	2
C*	0	0.9944	0.0852	0
	1	0.0056	0.9148	0.0934
	2	0	0	0.9066

Haplotype H12, Frequency 0.023

		H		
		0	1	2
C*	0	0.9936	0.2096	0
	1	0.0064	0.79	0.2586
	2	0	0.0004	0.7414

1%

Haplotype H22, Frequency 0.124

		H		
		0	1	2
C*	0	0.9497	0.1394	0.0036
	1	0.0503	0.8599	0.2266
	2	0	0.0007	0.7698

Haplotype H2, Frequency 0.053

		H		
		0	1	2
C*	0	0.9941	0.1672	0
	1	0.0059	0.8328	0.1833
	2	0	0	0.8167

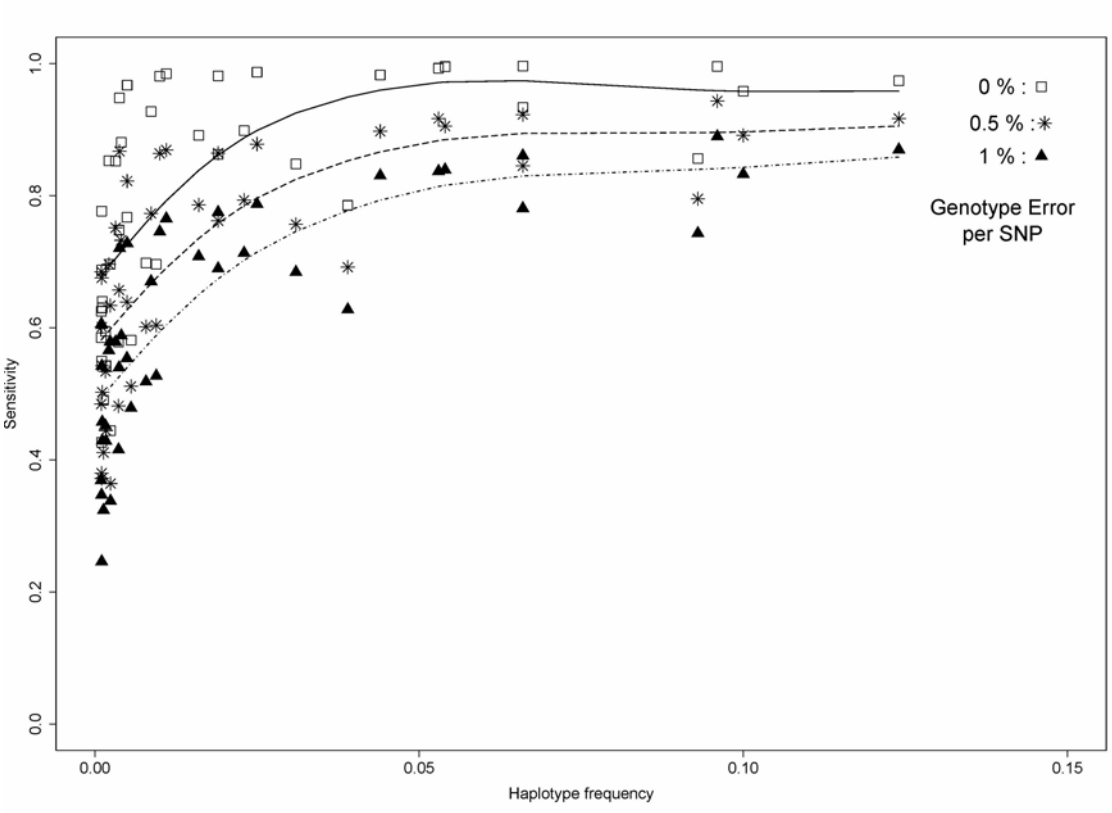
Haplotype H12, Frequency 0.023

		H		
		0	1	2
C*	0	0.9929	0.2902	0.0345
	1	0.0071	0.7091	0.3276
	2	0	0.0007	0.6379

4.2.1.2. Estimated sensitivity and specificity for haplotype misclassification

A summary using the sensitivity and specificity is provided by depicting these in Figure 18 for all *APM1* haplotypes illustrating the dependence on haplotype frequencies. The sensitivity is high for the common haplotypes and for many rare haplotypes, but is substantially decreased down to 50% for some rare haplotypes (Figure 18 a). It can further be seen that the genotype error adds to a decrease of the sensitivity in a rather "parallel" fashion, that is rather independent from the haplotype frequency. The sensitivity is down to 40% for some haplotypes with a genotype error of 1%. The specificity is reduced for the common haplotypes, but is 100% for most of the rare haplotypes (Figure 18 b). It also decreases with added genotype error but never below 95%.

(a)



(b)

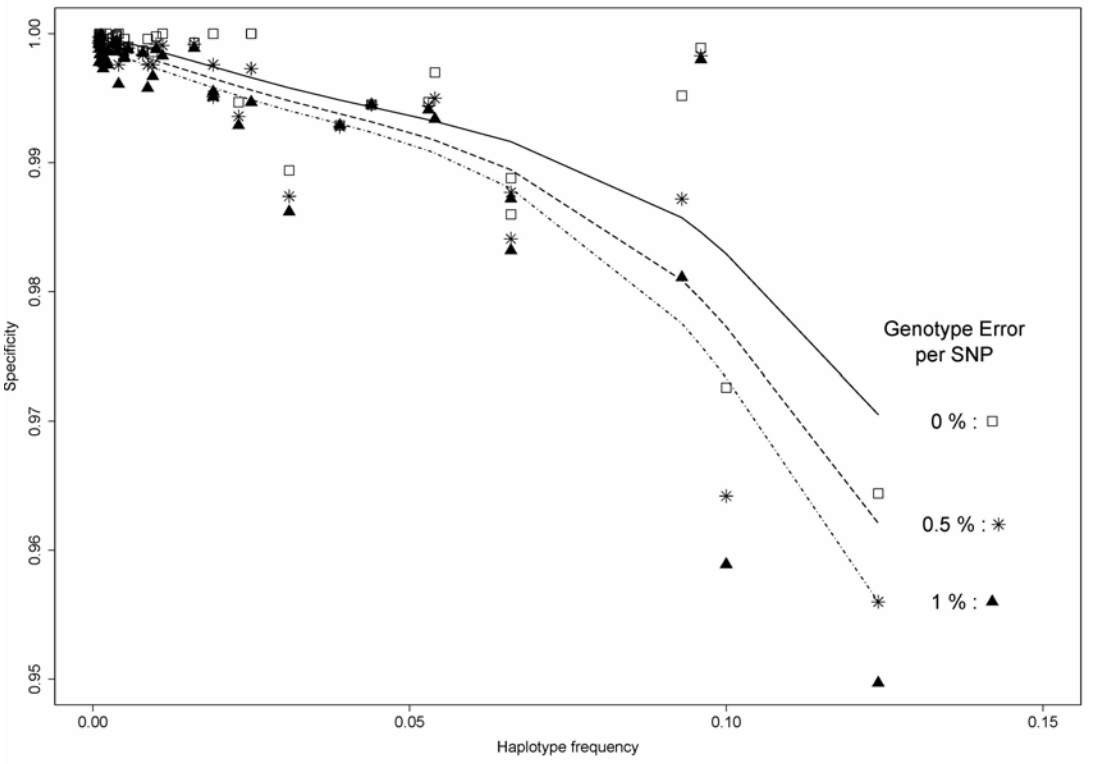


Figure 18: *Dependency of sensitivity (panel a) and specificity (panel b) on haplotype frequency from APM1 gene haplotypes for 0, 0.5 and 1% genotype error.*

The overall error rate of reconstructed haplotypes increases from 0.0667 in the case of no genotype error to 0.2015 and 0.3136 in the case of 0.5 or 1% genotype error, respectively, again illustrating the substantial influence from genotype error as compared to pure haplotype reconstruction error. The overall discrepancy, however, is rather small, but also increases for increasing genotype error: 0.0199 for 0%, 0.0738 for 0.5% and 0.1253 for 1% genotype error. Figure 19 depicts the haplotype-specific discrepancy again sorted for the haplotype frequencies. It can be seen that the discrepancy per SNP is small throughout: it does not exceed 0.001 in the case of no genotype error and reaches its maximum at 0.006 for 1% genotype error.

It should be noted that haplotype reconstruction error and genotype error not only evoke wrongly assigned haplotypes, which is grasped by the misclassification matrix, but they also "create" new haplotypes. Primarily in the case of genotype error, haplotypes are obtained from reconstruction, that are not present in reality. The percentage of wrongly created haplotypes increased from 0.395% for no genotype error, to 6.30% and 11.50% for 0.5% and 1% genotype error, respectively. The frequencies of these wrongly created haplotypes, however, did not exceed 0.25%.

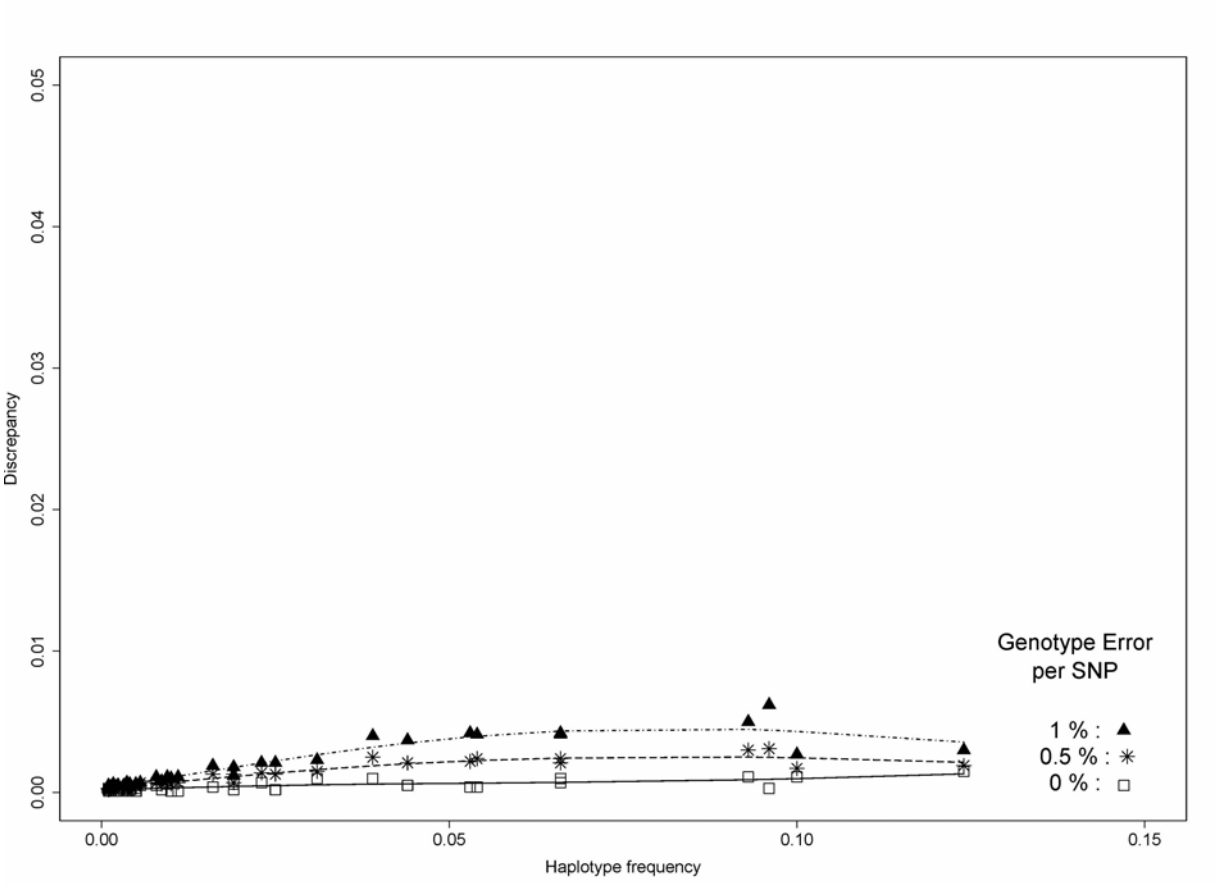


Figure 19: *Dependency of discrepancy D_m on haplotype frequency from APM1 gene haplotypes for 0, 0.5 and 1% genotype error.*

4.2.2. Bias in estimates and performance of MC-SIMEX

Simulation results of haplotype association analysis showing the bias in the estimate from the misclassification as well as the performance of the MC-SIMEX to correct this bias are summarized in Table 15. While the estimates of the naive model, which is the usual analysis ignoring the misclassification, clearly underestimates the true haplotype effect for all dominant models (Table 15 (a)), the MC-SIMEX estimates approximate the true estimate very well. For low misclassification (Π_{low}), all three extrapolation functions show good results, having estimates closest to the true model for the quadratic and loglinear function. In the case of high misclassification (Π_{high}), there is still clear underestimation when using the linear extrapolation function for the MC-SIMEX estimator, showing the best results for the loglinear extrapolation function. The coverage rate, a measure of how well the type I error is preserved, reaches the 95%-threshold, that has been expected, for all true dominant models. In almost all cases based on misclassification matrix Π_{low}, the MC-SIMEX estimators resulting from the quadratic or loglinear extrapolation function also reach this threshold; for matrix Π_{high}, the coverage is highest for the loglinear function. In this case, the naive model clearly deviates from the true model, also with respect to the coverage. Almost the same results could be observed for β_1 and β_2 of the additive model (Table 15 (b)). Due to the smaller number of homozygote rare haplotype pairs, the standard deviation of the β_2 estimates are higher and thus the corresponding coverage. None of the results, if based on a dominant or additive genetic effect, depend on haplotype frequency.

The relative bias, defined as $\frac{\hat{\beta}_{naive} - \beta}{\beta} * 100\%$, ranges between -9.9 and -12.4 % for Π_{low} and -33.8 and -39.4 % for Π_{high} for the dominant model. In the case of an additive model, the relative bias for β_1 ranges between -8.5 and -10.1 % (Π_{low}) and -31.0 and -33.1 % (Π_{high}). For β_2 it ranges between -1.2 and -4.7 % (Π_{low}) and -11.5 and -17.0 % (Π_{high}).

In summary, the absolute deviation from the true estimate is higher for higher effect sizes compared to the smaller effect sizes (0.5 vs. 0.05). The relative deviation, however, remains about the same with respect to the effect size, but is substantially higher for higher misclassification as expected. The MC-SIMEX corrected estimate removes most of the bias from haplotype misclassification, especially for the quadratic and loglinear extrapolation function. For some cases, though, the mean of the loglinear MC-SIMEX estimate slightly overestimates the true estimate, while the quadratic function shows more conservative results.

Table 15: *Performance of MC-SIMEX in a linear regression analysis for haplotype association for a specific risk haplotype h_R with given haplotype probabilities and misclassification schemes A and B. Linear regression analyses were performed ignoring the haplotype error ("naive model"), and accounting for the error fitting a linear, quadratic, or loglinear function to the SIMEX-simulated estimates assuming a dominant (Table (a)) and additive (Table (b)) genetic model.*

(a)

Estimator	Misclassification Matrix Π_{low}		Misclassification Matrix Π_{high}	
	Mean($\hat{\beta}$)	Coverage rate	Mean($\hat{\beta}$)	Coverage rate
$\beta_{dom} = 0.5$, $\sigma^2 = 0.4$, $f_{h_R} = 0.15$				
True model	0.500291	0.950	0.500291	0.950
Naive Model	0.448636	0.600	0.311041	0.000
Simex (linear)	0.491278	0.925	0.394515	0.215
Simex (quad)	0.498974	0.940	0.469148	0.860
Simex (loglin)	0.500178	0.955	0.499947	0.900
$\beta_{dom} = 0.5$, $\sigma^2 = 0.4$, $f_{h_R} = 0.3$				
True model	0.501029	0.965	0.501029	0.965
Naive Model	0.444911	0.465	0.324843	0.000
Simex (linear)	0.486050	0.920	0.405588	0.225
Simex (quad)	0.496692	0.955	0.467759	0.820
Simex (loglin)	0.494501	0.950	0.488627	0.915
$\beta_{dom} = 0.05$, $\sigma^2 = 0.4$, $f_{h_R} = 0.15$				
True model	0.050291	0.950	0.050291	0.950
Naive Model	0.045312	0.930	0.030478	0.855
Simex (linear)	0.049565	0.935	0.038807	0.880
Simex (quad)	0.050526	0.940	0.045819	0.925
Simex (loglin)	0.050524	0.940	0.050617	0.910
$\beta_{dom} = 0.05$, $\sigma^2 = 0.4$, $f_{h_R} = 0.3$				
True model	0.051029	0.965	0.051029	0.965
Naive Model	0.044715	0.950	0.033794	0.900
Simex (linear)	0.048905	0.950	0.042219	0.900
Simex (quad)	0.049712	0.960	0.048225	0.910
Simex (loglin)	0.049982	0.950	0.051700	0.935

(b)

Estimator		Misclassification Matrix Π_{low}		Misclassification Matrix Π_{high}	
		Mean($\hat{\beta}$)	Coverage rate	Mean($\hat{\beta}$)	Coverage rate
$\beta_{add} = 0.5,\ \sigma^2 = 0.4,\ f_{h_R} = 0.15$					
True model	β_1	0.500050	0.940	0.499852	0.930
	β_2	1.005247	0.935	0.995633	0.955
Naive Model	β_1	0.449460	0.655	0.334322	0.005
	β_2	0.993179	0.945	0.840234	0.510
Simex (linear)	β_1	0.490812	0.910	0.417977	0.400
	β_2	1.006970	0.940	0.984643	0.960
Simex (quad)	β_1	0.497946	0.930	0.480312	0.890
	β_2	1.010417	0.940	0.992940	0.900
Simex (loglin)	β_1	0.499131	0.925	0.504503	0.900
	β_2	1.007353	0.940	1.049798	0.920
$\beta_{add} = 0.5,\ \sigma^2 = 0.4,\ f_{h_R} = 0.3$					
True model	β_1	0.499882	0.925	0.498705	0.960
	β_2	0.994083	0.980	0.999623	0.965
Naive Model	β_1	0.452785	0.620	0.342696	0.000
	β_2	0.951016	0.855	0.829435	0.000
Simex (linear)	β_1	0.488278	0.910	0.421349	0.000
	β_2	0.980813	0.960	0.942516	0.260
Simex (quad)	β_1	0.496652	0.945	0.472260	0.465
	β_2	0.991524	0.975	0.966305	0.585
Simex (loglin)	β_1	0.494283	0.940	0.487507	0.860
	β_2	0.982675	0.965	0.980507	0.900
$\beta_{add} = 0.05,\ \sigma^2 = 0.4,\ f_{h_R} = 0.15$					
True model	β_1	0.050050	0.940	0.049852	0.930
	β_2	0.105247	0.935	0.095632	0.955
Naive Model	β_1	0.044999	0.930	0.034027	0.880
	β_2	0.109134	0.935	0.084642	0.945
Simex (linear)	β_1	0.049073	0.930	0.042545	0.905
	β_2	0.110251	0.940	0.098878	0.940
Simex (quad)	β_1	0.049693	0.925	0.048512	0.925
	β_2	0.110567	0.940	0.100803	0.925
Simex (loglin)	β_1	0.050071	0.940	0.054005	0.935
	β_2	0.110512	0.940	0.108009	0.935
$\beta_{add} = 0.05,\ \sigma^2 = 0.4,\ f_{h_R} = 0.3$					
True model	β_1	0.049882	0.925	0.049882	0.925
	β_2	0.094083	0.980	0.094083	0.980
Naive Model	β_1	0.045636	0.940	0.034407	0.910
	β_2	0.089645	0.970	0.079686	0.965
Simex (linear)	β_1	0.049268	0.940	0.042218	0.915
	β_2	0.092564	0.970	0.090419	0.955
Simex (quad)	β_1	0.050070	0.950	0.046998	0.940
	β_2	0.094382	0.980	0.094288	0.935
Simex (loglin)	β_1	0.050100	0.945	0.050959	0.960
	β_2	0.093045	0.975	0.095271	0.965

4.2.3. *APM1* data example: MC-SIMEX-corrected haplotype association estimates

Applying the MC-SIMEX to correct the association of the *APM1* haplotypes with plasma adiponectin concentrations in the SAPHIR study for pure haplotype misclassification from reconstruction yielded higher estimates for almost all haplotypes. Also, correcting for the genotype error additional to the reconstruction error further increased the resulting estimates. Figure 20 summarizes these results. Generally, the absolute correction of estimates, computed as the difference between β_{simex} and β_{naive}, was larger for higher effect sizes, as already seen in the simulations, while the relative bias depended neither on effect size nor on haplotype frequency. Correcting only for reconstruction error, the relative bias of the naive estimate with respect to the MC-SIMEX corrected estimate was as high as -21%, while it was up to -39% and -48% for 0.5 and 1% genotype error, respectively.

For haplotype H1, for example, with rather high association with plasma adiponectin, β_{naive} was 0.2196 and $\beta_{simex,1\%}$= 0.2813 (see Figure 20). As Plasma-Adiponectin entered the linear regression model as log(adiponectin+1) to yield a normal distribution, this corresponded to a difference of 13.75 µg/dL for women with at least one copy of the H1 haplotype compared to women with the reference (no copies of H1) without accounting for the misclassification, while the difference was 14.64 for women when accounting for the misclassification (assuming 1% genotype error). For men, the difference was 8.76 without and 9.31 with accounting for misclassification. This corresponds to an underestimation by 0.89 µg/dL for women and 0.55 µg/dL for men, respectively, when misclassification is not accounted for. The relative bias was -22%. For a haplotype with rather small association effect, e.g. haplotype H11, the absolute underestimation when ignoring the misclassification was 0.019 for women and 0.012 for men, with a relative bias of -4.3%.

Results for a 2df-model depicting the effect for subjects heterozygous (β_1) and for subjects homozygous (β_2) of the respective haplotype compared to the reference are also shown in Figure 20. This model was only calculated for haplotypes H16, H26 and H15, where more than 10 subjects were homozygous for this haplotype. For comparison, the estimate based on the expected value of haplotypes (β_{H*}), was also plotted. If the additivity assumption truly held, β_2 would be two times β_1 and β_{H*} would be a possibility to account only for the error from reconstruction. In this data, the additivity assumption doesn't necessarily hold for haplotypes H16 and H15, but could be assumed for haplotype H26. In this case β_{H*} =0.0385

underestimates the MC-SIMEX corrected estimate for pure reconstruction error ($\beta_{1,SIMEX\ 0\%}$ = 0.0417) only slightly, whereas it deviates slightly more for haplotypes H16 and H15. It is obvious for all three haplotypes, that misclassification through genotype error cannot be corrected for by simply using the haplotype trend regression estimates β_{H*}.

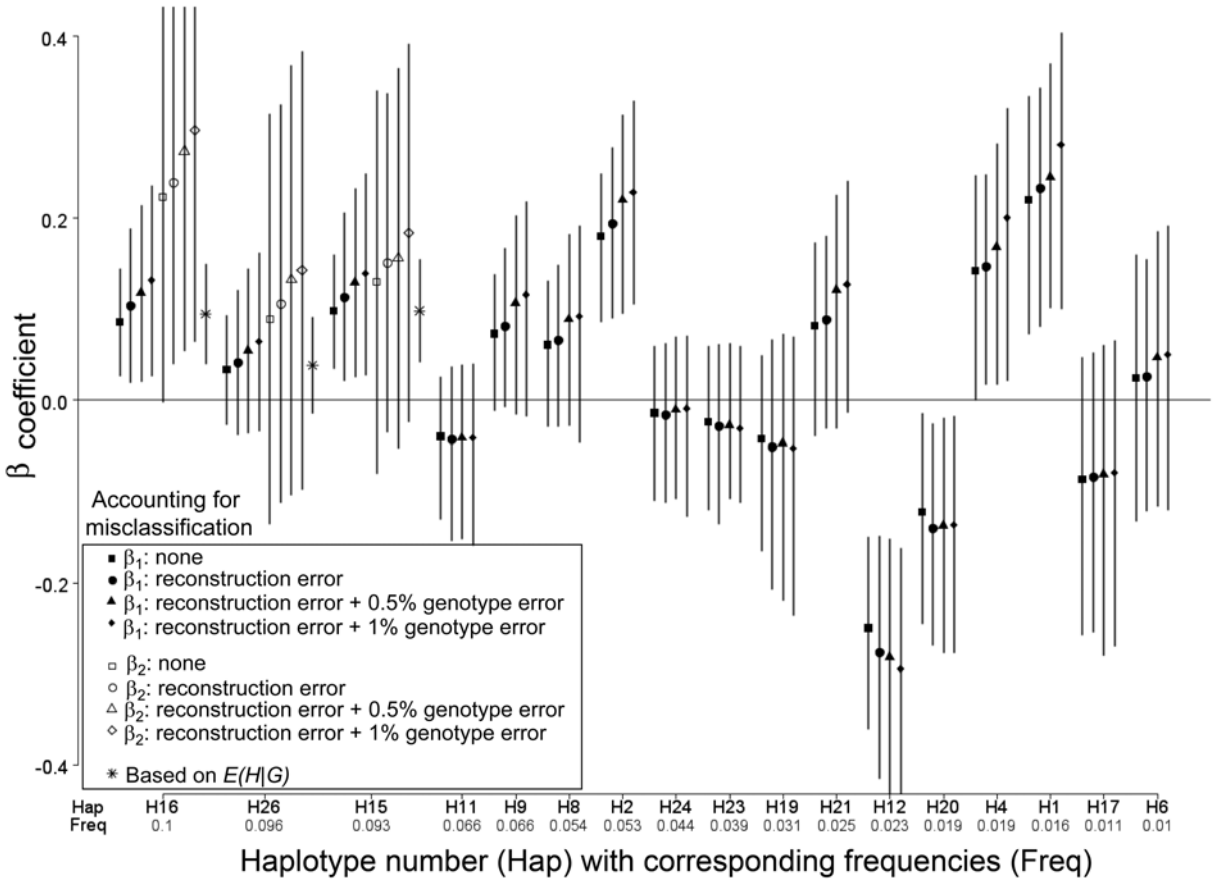


Figure 20: *Results for 18 haplotypes from the APM1 gene with frequency>1%: β-Estimators for a linear regression model on log(adiponectin +1) adjusted for age, sex and all other haplotypes with the most frequent one as reference, assuming no error at all (naive estimator), only reconstruction error, and reconstruction with 0.5% and 1% genotype error. For the three most frequent haplotypes H16, H26 and H15, estimates are given for one (β_1) and two copies of each haplotype (β_2), as well as based on the expected number of haplotypes (β_{H*}). For all other haplotypes, β_1 is the effect for having at least one copy of the haplotype.*

5. Discussion

5.1. Summary of main results

It was the objective of this work to quantify and characterize the misclassification of genetic variants and its impact on genetic association studies. The first part addressed the misclassification in genotypes that results from the genotyping process as it can be expected for genetic association analysis in the most general form. Biologically reasonable error models were considered to find the most parsimonious model describing the misclassification problem without loosing information. To address this objective, a representative set of routinely duplicate genotyped SNPs has been selected. Evaluation of genotyping misclassification probabilities resulted in estimates of 0.001 and below.

The data suggested the allelic drop out model to fit best among the considered misclassification models. The symmetric model was found to fit reasonably well while being at the same time more parsimonious. Furthermore, evidence was provided against further restrictions.

The bias of such a genotype misclassification, as shown in a re-analysis of the association of SNPs of the *APM1* gene on plasma adiponectin concentrations by applying the MC-SIMEX method to correct for misclassification, was up to 15% under an extreme genotype error setting, but almost non-existing for the estimated und thus more realistic size of genotyping error.

The second objective concerned the amount and structure of uncertainties in haplotypes induced by statistical reconstruction. A classification of the various error measures was provided. Sensitivity and specificity, well-known measures from other areas of biomedical research, were introduced to the context of haplotypes and an analytical computational approach was presented. The quantity of the various error measures and their dependencies upon haplotype frequencies were illustrated in a systematic way, including minor allele frequency, correlation, number of loci, and ambiguity fraction. The results emphasize the dependence of the haplotype reconstruction error on the specific situation and the importance of haplotype-specific error measures.

Finally, the combined impact of genotyping error and haplotype reconstruction error on haplotype association estimates was of interest. The simulations to quantify haplotype misclassification including potential genotype error showed substantial misclassification for some haplotypes by pure reconstruction error and aggravation by a genotype error of 0.5% to

1%, that might be in a realistic degree. For haplotype association analysis, simulations were performed showing a bias of up to 40% for high misclassification, and still 12% for small misclassification in the naïve model ignoring this misclassification. The MC-SIMEX corrected estimates were shown to be very close to the true effects with coverage rates reaching the 95%-threshold. Within the real data example on *APM1*, MC-SIMEX corrected estimates yielded higher estimates for almost all haplotypes and slightly extended confidence intervals. In this work, the MC-SIMEX approach was introduced to genotype and haplotype association analysis for the first time to correct for misclassification.

5.2. Quantification of misclassification

5.2.1. Misclassification of SNP genotypes

So far, the genotyping error was never before estimated in a set of representative epidemiological samples as it is encountered in routine association analyses. To obtain an estimate of the genotyping error in the epidemiological praxis, a sufficiently large data set with routinely performed repeated genotypes had to be collected, which presented a great challenge. Assumptions or restrictions to the error model had to be made to obtain an identifiable 3x3 genotype misclassification problem with double genotypes. Methods requiring more than two genotype repetitions could not be applied due to adaption on routine epidemiological data. Furthermore, it was one of the main objectives to estimate an unrestricted error model that is as general as possible. Restrictions to the error model would also have omitted the possibility to explore the misclassification model fit. Therefore, the same misclassification for all SNPs was assumed. This assumption was supported by the data as the discordance did not depend upon the MAF (see Figure 10b). This assumption is also practical when the interest lies in the overall error in the genotypes across a full set of SNPs rather than in the error of a specific SNP. It also enabled the estimation of the genotyping error under the most general error model without any restrictions, while the literature covers rather restricted error models. The error models that were discussed so far in the literature and that were primarily used for simulations, are summarized in Table 16.

The results of this investigation suggested the allelic drop-out model as appropriate. The symmetrical model was also shown to fit reasonably well while being at the same time more parsimonious involving three instead of six parameters, and provided evidence against further restrictions. It should be noted that estimation of all 6 parameters was not as robust as

desirable when excluding observations, which is most likely due to the very small size of the error giving rise to only 210 discordant genotype pairs despite the large sample size. Nevertheless, the fact that there were discordant pairs with both opposite homozygous genotypes suggests that the "zero corner model" is not appropriate. The allele-independent model, which is most widely applied in the literature so far, can be considered as most appealing for all practical purposes due to its simplicity, its allowing for "non-zero corners", and the robustness of estimates. Although this model was not supported in this investigation, in cases of very small genotype error as it was detected in this data, the particular choice of a specific error model is not a concern.

Overall, very small misclassification probabilities were found. This investigation thus underscores the validity of SNP genotypes under situations comparable to this study and may in fact explain – to some extent – the success of SNP association studies. It should be noted, however, that such a small error cannot be expected when relaxing quality control procedures or when using a less established new genotyping technology such as genome-wide SNP chips.

5.2.2. Misclassification due to haplotype reconstruction error

The simulations on haplotype uncertainty due to haplotype reconstruction alone resulted in the following: The discrepancies between the observed and true haplotype frequencies were observed to be less than 0.005, which supports the observation of Fallin and Schork [Fallin and Schork, 2000], who also described a small discrepancy. However, the observation of small discrepancies can not be transferred to individuals' haplotype error, for example the error rate.

The overall error rate is the most reported error measure [Adkins, 2004;Niu et al., 2002;Stephens et al., 2001;Stephens and Donnelly, 2003;Xu et al., 2004] . It depends heavily upon the specific setting: The error rate was generally increasing with increasing number of loci, increasing minor allele frequency, decreasing correlation between the alleles, and increasing ambiguity fraction. While the error rate was small in some real data examples (e.g. *IL-18*, *INS* with error rate <1%), it was substantial in others (*CAPN10*, *TNFA* with error rate up to 12%). The error rate is useful to provide a general picture of haplotype error for a gene. However, an investigator is usually interested in a specific risk haplotype and in how to interpret this haplotype's association estimate. Then the question arises whether this specific haplotype is reconstructed with great error, and the error rate averaging across all haplotypes is not of much help.

Table 16: *Overview of genotype misclassification models and estimated error size in the literature*

Error model	*Description*	*Parameterization*	*Literature Reference*	*Estimated error size*
"*General*"	No Restrictions	π_{ij} = P(observed genotype = i | true genotype = j), i,j=0,1,2	[Hao and Wang, 2004;Kang et al., 2004b;Moskvina and Schmidt, 2006]	-
"*Allelic drop out*"	One allele signal vanishes beneath white noise and thus P(hom→het)>P(het→hom);	$\pi_{01} > \pi_{10}$ and $\pi_{21} > \pi_{12}$	[Mitchell et al., 2003;Morris and Kaplan, 2004]	-
"*Zero corner*"	Hom major* never misclassified as hom minor and vice versa	$\pi_{20} = 0$, $\pi_{02} = 0$	[Morris and Kaplan, 2004]	-
"*Symmetrical*"	Misclassification does not differ for hom major or hom minor	$\pi_{10} = \pi_{12}$, $\pi_{01} = \pi_{21}$ and $\pi_{20} = \pi_{02}$	-	-
"*Hom-het*"	Zero corner and symmetry as described above	P(hom→het)=:ν, P(het→hom)=:μ [$]	[Gordon et al., 2002;Leal, 2005;Seaman and Holmans, 2005]	-
"*Directed error*"	Error described per allele	P(A→a)= :μ, P(a→A)=:ν [&]	[Akey et al., 2001b;Gordon et al., 2002;Gordon and Ott, 2001;Leal, 2005;Moskvina and Schmidt, 2006;Ritchie et al., 2003;Zou et al., 2003]	-
"*Allele-independent*" (also "*Stochastic error*")	Error described per allele, assumed as independent from allele Special case of symmetrical model; related to zero corner model, if e is small as then e^2 is close to zero.	P(A→a)=P(a→A):=ε (see also Table 1c)	[Akey et al., 2001b;Becker et al., 2006;Govindarajulu et al., 2006;Hao and Wang, 2004;Kirk and Cardon, 2002;Moskvina and Schmidt, 2006;Wong et al., 2004]	ε = 0.0074 [Wong et al., 2004] one study with 1027 subjects genotyped twice to estimate error
"*Uniform error*"	Uniform misclassification probabilities	$\pi_{ij} = e$ for $i \neq j$, i,j=0,1,2	[Cox and Kraft, 2006;Kang et al., 2004b;Lincoln and Lander, 1992;Liu et al., 2006;Mitchell et al., 2003;Rice and Holmans, 2003;Sobel et al., 2002]	e = 0.014 from repeated typing [Lincoln and Lander, 1992]

* Hom minor and hom major: short cut for homozygous of the minor or major allele, respectively.

$ P(hom→het): the transition probability for true homozygous genotype (no matter if minor or major) misclassified as heterozygous, P(het→hom) vice versa.

& P(A→a): the transition probability for true major allele A misclassified as the minor allele a, P(a→A) vice versa

A known haplotype-specific measure for haplotype reconstruction error is R_m^2, which indicates the proportion of haplotype variance explained by the genotypes. It captures the haplotype-specific error $H_m \rightarrow H_m^*$, which is the error from using the individual's expected instead of the true number of copies of a haplotype. This is a very complex error model, as it moves from the discrete space {0,1,2} to the continuous space [0,2] with the distribution of H_m^* being three-modal at 0,1, and 2. This error measure can only be applied when using the expected number of haplotypes H_m^* as explaining variables in the haplotype association analysis.

An alternative is a haplotype-specific measure for the error $H_m \rightarrow C_m^*$, that is the error from using the individual's most likely number of copies of the haplotype instead of the true number of copies. This is an error model from the discrete space {0,1,2} into the discrete space {0,1,2} and thus a classical misclassification problem, which is represented by the 3x3 misclassification matrix. This is very appealing as the full concept of misclassification is then available. When the misclassification matrix is known, methods are available to account for the error, e.g. by means of the matrix method [Morrissey and Spiegelman, 1999] or the MC-SIMEX [Kuchenhoff et al., 2006].

As new notions of haplotype-specific error measures, the sensitivity and the specificity were introduced into the context of haplotypes. Both measures complement the R^2 parameter and differentiate between two reasons for haplotype reconstruction error:

Firstly, the specificity is an issue for common haplotypes: If the specificity is reduced, it is reduced rather for a common haplotype (Figure 16 and Figure 17). This is plausible due to the fact that if any haplotype is misclassified, it is rather misclassified as a common haplotype by pure chance. Therefore, a common haplotype is more likely falsely assigned than a rare haplotype.

Secondly, the sensitivity is an issue for rare haplotypes: If the sensitivity is reduced, it is reduced more likely for a rare haplotype. For example, the rather low sensitivity of the haplotype 101 of *MCP1* (Figure 16) was due to the fact, that this haplotype most likely paired with the most common haplotype 000 given Hardy-Weinberg equilibrium (haplotype pair 101/000) and that the alternative haplotype pair 001/100 contained two rather frequent haplotypes (001 and 100 with frequencies 5.4% and 17.5%). Thus, the haplotype pair 101/000 would often be falsely reconstructed as 001/100. Generally speaking, the haplotype pair containing a rare haplotype - and thus the rare haplotype itself - is more likely falsely classified. On the other side, there are also rare haplotypes that are perfectly reconstructed, which occurs when there is no likely alternative haplotype pair. For example, the haplotype

0010101 of *CAPN10* (Figure 17) showed almost 100% sensitivity: Besides the haplotype pair consisting of this rare haplotype and the most common haplotype, 0010101/0000000, an alternative would have been 0010001/0000100 or 0010000/0000101. However, for both alternatives, either of the two haplotypes did not exist with frequency >1%. Thus the probability of such a pair was negligible, the pair 0010101/0000000 was assigned with great certainty, and the rare haplotype 0010101 was therefore very well reconstructed.

Sensitivity and specificity were also shown to completely describe the misclassification matrix (Table 10) and thus provide the prerequisite for methods accounting for misclassification.

The quantity of reconstruction error is hard to predict intuitively as the reconstruction depends on the full constellation of the other haplotypes. To better judge whether the haplotype association estimate is biased due to substantial reconstruction error, looking at the haplotype-specific error measures would greatly enhance the knowledge about the reliability of haplotypes and respective association estimates. Therefore, a graphical tool was developed to comprehensively display the haplotype-specific error measures R_m^2, sensitivity or specificity (see Appendix).

The analytical derivations of error rate, sensitivity and specificity complement the computational formula of R_m^2 [Stram et al., 2003b]. The simulations validated the analytical approach also comparing EM- versus PHASE-reconstruction. It should be noted that the error measures in the simulations included the sampling error and were thus slightly higher than the analytically derived measures, but the difference was not substantial due to sufficient sample size. Comparing EM- with the PHASE-reconstruction, both methods were found to work equally well when applying *real data scenarios*. The *abstract scenarios*, while being useful to make extreme examples and to understand mechanisms, included situations such as the no-LD scenario under which no haplotypes should be reconstructed in the first place.

5.2.3. Misclassification due to genotyping error and haplotype reconstruction error combined

To enable the application of misclassification correction methods such as the MC-SIMEX on haplotype association, the matrix of misclassification probabilities has to be known or estimated, for example with the resampling approach presented in this work. With sensitivity and specificity, the full misclassification matrix is determined for dichotomous variables, e.g.

when assuming a dominant or recessive effect on haplotypes. Sensitivity and specificity are also sufficient to describe the misclassification matrix, if only pure reconstruction error (i.e. no genotype error) applies. It is easily seen, that genotype error adds to these corners of the misclassification matrix. Thus, the misclassification matrix was extended to a full 3x3-matrix for model-free inheritance and/or genotype error. A simulation approach was proposed to estimate misclassification probabilities, sensitivity and specificity. This approach was exemplified for a real data set of the *APM1* gene with complex haplotype structure. Double sampling or molecular haplotyping might be an alternative to assess haplotype misclassification probabilities. Levenstien et al. [Levenstien et al., 2006] presented a method which uses molecular haplotypes on a subset of individuals to estimate haplotype misclassification and account the Likelihood Ratio test for it in the setting of case-control studies. However, due to the absence of high throughput procedures for molecular haplotyping, this method is too time- and money-consuming to be a practical approach. Furthermore, even laboratorily assessed haplotypes are subject to error. Thus, due to the lack of a gold-standard procedure, an estimation of expected haplotype misclassification was provided based on the frequencies of observed haplotypes and assumed genotype error rate. Since haplotype specific discrepancies were found to be very small throughout, this approach seemed to be suitable. The exact calculation of haplotype misclassification matrices isn't possible anyway, since it always involves genotyping error, which has to be assumed or derived from other studies. Here, genotype error was assumed to be allele-independent, since it has been shown to be the most robust while being the most parsimonious model.

Genotype error added to the misclassification from statistical haplotype reconstruction rather independently from haplotype frequency. In contrast, haplotype reconstruction error alone was previously shown to depend upon haplotype frequency. The observed dimension of misclassification can be a problem, when specific individuals are to be picked based on their haplotypes for further follow-up, for example for further investigations or in-depth studies. Thus, deriving the potential of the respective haplotype to imply misclassification can be an intriguing tool to ascertain that the correct subjects are picked for these expensive studies.

5.3. Impact of misclassification on genetic association

5.3.1. Impact of genotype misclassification on association estimates

To investigate the impact of SNP genotype misclassification on association estimates, the association of 13 *APM1* gene SNPs on plasma adiponectin levels was reanalyzed applying the MC-SIMEX correction method. The results pinpointed only marginal bias on association estimates induced by an error as estimated by the repeated genotype data. However, it was illustrated that increased genotype error would decrease association estimates. The extreme genotyping error scenario showed a change of estimates after correction of up to 15% indicating that relaxed quality control or less established genotyping methods would imply substantial impact. The change in the estimate was away from the null, thus correcting for the bias towards the null in the uncorrected estimate, which is according to the statistical theory [Carroll et al., 2006]. In such cases of high genotyping errors, the MC-SIMEX approach can effectively remove this bias induced by this misclassification. As the MC-SIMEX can handle a wide range of error and association models also for this trichotomous variable situation and allows adjustment for other covariates, it can be recommended for utilization in future genetic association studies. It should be noted, though, that in cases with extremely low error rate, as observed in the duplicate genotype data presented in this study, methods such as the MC-SIMEX are probably not needed.

5.3.2. Impact of haplotype misclassification on association estimates

Since haplotype reconstruction always involves SNP genotypes that are potentially error-prone, the evaluation of haplotype misclassification impact on association estimates should always include potential genotype error.

The performance of the MC-SIMEX method on haplotype association analyses was estimated using full 3x3 misclassification matrices. Error-afflicted haplotypes with one risk haplotype were simulated given a certain parameter β on a normally distributed outcome variable. A bias of up to 40% for high misclassification and still 12% for small misclassification in the naïve model ignoring misclassification indicate that haplotype misclassification shouldn't be ignored in haplotype association analysis. This amount of bias is slightly higher than the maximum bias of -5.8%-32.3% for different haplotype blocks that have been found in

Govindarajulu et al. [Govindarajulu et al., 2006], accounting for 1% genotyping error. However, there was no reconstruction error assumed in their findings.

MC-SIMEX corrected estimates, especially using the loglinear extrapolation, were very close to the true effect with coverage rates reaching the 95%-threshold Thus, the MC-SIMEX has been shown to correct the bias that is involved with haplotype error very well.

To further show the impact of haplotype misclassification on realistic association estimates, the MC-SIMEX correction has been used on an association example on *APM1* gene with adiponectin plasma levels. The influence of the genotype error, accumulating across the number of SNPs, on the overall misclassification on *APM1* haplotypes was larger than the pure haplotype reconstruction error. Thus, haplotype reconstruction error alone affected association estimates only slightly. It has to be noted, that association analysis was only conducted for haplotypes with a frequency higher than 1%. Large declines in sensitivity by reconstruction error alone could only be observed for very rare haplotypes (Figure 18). Due to their rare occurrence, however, these badly reconstructed haplotypes shouldn't play a major role in haplotype association analysis.

Altogether, MC-SIMEX corrected estimates were higher for almost all haplotypes and confidence intervals were slightly extended. Slightly corrected estimates towards the null occurred only in the case of very small haplotype effects (e.g. H24 and H17) and might thus be due to adjustment by all other haplotypes. For all other haplotypes, comparing the SIMEX-corrected estimates to the naive estimates, there was a bias towards the null of up to almost -50%. For adiponectin, an underestimation by 0.89, as it was observed for haplotype H1, is already clinically relevant. However, if staying in the range of genotype error that might be expected for good genotyping laboratories, say 0.5% or even lower, the impact seems to be rather low and of minor relevance to the data presented here. This finding stresses the importance for high genotyping quality.

5.4. Strengths and limitations of this investigation

5.4.1. Issues regarding genotype misclassification

This investigation was based on the collection of a large representative set of epidemiological studies with double genotype data as it is encountered in the genetic epidemiological practice. This sample was not an experimental data set and laboratory personnel were not aware of this project at the time of genotyping. That is, misclassification probabilities estimated from this

routine epidemiological data are expected to be the same as encountered in data sets used for analysis with sufficient quality control, which revealed as a great strength of this investigation.

Furthermore, the approach can be applied in routine association analyses: After estimating genotyping from routinely performed double genotypes, the MC-SIMEX can be applied for most analyses models such as linear and logistic regression allowing for all kind of adjustment, all genetic effect models and most misclassification models. It was a further strength of this approach that an unrestricted error model could be applied, which was not reported previously.

It must be considered a limitation of this study, that double genotyping by the same genotyping method using the same DNA and aliquots does not enable to grasp all error sources. Furthermore, mismatch of DNA to subjects will be undetected. Thus, conclusions can only be made about some aspects of the genotyping error. Furthermore, the genotype model and association analyses was based on categorized genotypes instead of genotype probability scores, which is more sophisticated from a statistical methodological perspective. However, routine genetic association analyses in a candidate gene approach currently use the genotype categories and the epidemiological practice was the focus here. Finally, the impact of differential error has not been evaluated [Moskvina et al., 2006]. Thus, the conclusions about the high validity of SNP genotype association results can neither be transferred to differential error in case-control studies nor to more complex genetic variants such as microsatellite markers implying a higher dimensional misclassification problem and a more error-prone technology.

5.4.2. Issues regarding haplotype misclassification

The classification and systematic investigation of error measures is a useful guidance for researchers interested in haplotypes and haplotype association estimates. This was strengthened by applying both analytical and simulation approaches for numerous scenarios, by exemplifying the measures to real data, and by utilizing the two basic and commonly used reconstruction methods, the EM algorithm and PHASE. Finally, this is the first work investigating the sensitivity and specificity of haplotype reconstruction and illustrating their impact on haplotype association analyses.

It might be considered a limitation that reconstructed haplotype frequencies from real data were used as “true” haplotype frequencies for the *real data scenarios*. However, this is

an excellent procedure to yield near-realistic haplotype distributions. The discrepancy was rather small. Therefore, reconstructed haplotype frequencies could be assumed to approximate the true frequencies fairly well. Due to the lack of a gold-standard, the expected haplotype misclassification can only be estimated based on the frequencies of observed haplotypes. Levenstien at al. [Levenstien et al., 2006] presented a method which uses molecular haplotypes on a subset of individuals to estimate haplotype misclassification and account the Likelihood Ratio test for it in the setting of case-control studies. However, due to the absence of high throughput procedures for molecular haplotyping, this method is too time- and money-consuming in most cases. Furthermore, even laboratorily assessed haplotypes are subject to error and can thus also not be taken as a gold-standard procedure.

Simulations were restricted to haplotypes across 2 to 7 loci, while in practice there are up to 20 loci. The restriction was made for the sake of limiting the complexity hypothesizing that the general findings can be transferred to longer haplotypes. Finally, Hardy-Weinberg equilibrium (HWE) was assumed for the haplotype pairs and this investigation did not evaluate the impact of violation of this assumption as other work has already focused on this issue (e.g. [Niu et al., 2002]).

Errors in individually assigned haplotypes can be presented by two different concepts: Firstly comparing the true number of copies of individuals' haplotypes (H) with the expected number of copies ($E(H|G^*)$), which is an error model from the discrete trichotomous space $\{0,1,2\}$ to the continuous space $[0,2]$ and secondly, comparing the most likely haplotypes ("best guess haplotypes", C*) with the true haplotypes, which is a classical 3x3 misclassification problem $H \rightarrow C^*$. The former one has the advantage to implicitly correct for haplotype reconstruction error [Mensah et al., 2007]. Since numerous methods have been developed to simultaneously estimate haplotype probabilities together with association estimates, this argument seems to have a big relevance. These methods do not infer individual haplotypes but use the expected values of haplotypes given the observed genotypes in association analysis within a likelihood framework [Epstein and Satten, 2003;Lake et al., 2003;Spinka et al., 2005;Zaykin et al., 2002] or with estimating equations [Zhao et al., 2003].

While the uncertainty in haplotype reconstruction error is accounted for in these analyses, they do not incorporate genotype error. They are often limited to case-control studies, can not incorporate environmental variables or assume additive effects, which is often not the case [Heid et al., 2006]. Individually inferred haplotypes, on the contrary, can easily be incorporated into generalized linear models (GLM), which provides wide flexibility in the modeling of underlying inheritance assumptions, the study type, the type of outcome variable

and gene-environment interactions. Due to the ease of computation in each standard statistical software, this method is quite popular in practice. Furthermore, if some haplotypes can be expected to be inferred correctly or with only small error with respect to sensitivity and specificity as in the example of the *CAPN10* haplotype mentioned, analyses based on inferred haplotypes is unbiased and is more powerful than the analysis based on expected haplotype probabilities [Little R.J.A and Rubin D.B., 2002].

However, association estimates can be biased substantially, if high haplotype misclassification is involved [Kraft et al., 2005;Schaid, 2004]. In these cases, information of the misclassification probabilities can be used to correct association estimates (e.g. using the MC-SIMEX method [Kuchenhoff et al., 2006]), and still stay in the flexible GLM framework.

5.5. Conclusions and Outlook

Genotyping error per SNP in a high quality laboratory with experienced personnel and an established genotyping method such as MALDI-TOF MS has been found to be small and negligible for SNP association studies. The approach to estimate genotyping error based on routine duplicate data has been shown to be applicable on general error models without parameter restrictions. It can be concluded that the MC-SIMEX method to account for genotype misclassification is suitable to SNP association analysis and can be utilized in cases, when the error in genotypes is more substantial e.g. for new genotyping technologies.

Furthermore, a classification and systematic quantification of haplotype reconstruction error measures has been provided. In certain situations, error measures like the error rate or discrepancy that sum over all haplotypes, are not sufficient to describe the error in certain haplotypes. These results underscore the importance of haplotype-specific error measures. The concept of sensitivity and specificity that was introduced to the context of haplotypes is particularly useful as it is well-known to life sciences researchers and thus easily communicated. An analytical computational approach was provided as well as a graphical tool for a summary presentation. Sensitivity and specificity can be quantified next to haplotype frequencies and haplotype association estimates to provide a sense of certainty into the haplotype reconstruction, especially if interest lies in one specific risk haplotype. Then, the misclassification matrix is known providing the necessary prerequisite for utilizing methods to account for misclassification.

An extension to a full 3x3 misclassification matrix was provided allowing for the inclusion of genotyping error, which can add substantially to the pure reconstruction error.

For haplotype association analyses, the MC-SIMEX was presented as an efficient method to calculate corrected association estimates, especially in the case of high genotype error. If a small genotype error rate can be assumed, however, the impact on haplotype association estimates is rather moderate, which underscores the validity of haplotype association analysis.

Furthermore, the knowledge of haplotype misclassification size and pattern can also be useful to researchers, which are interested in specific haplotype assignments beyond association analysis: e.g. haplotype assignments might be used to put up phylogenetic trees or to select certain persons for further studies. For these kinds of application, it has been shown, that specific haplotype assignments might generally be a concern for rare haplotypes if high genotype error can be expected. Therefore, a high quality genotyping, sufficient quality control and calculation of misclassification probabilities are necessary steps to minimize waste of resources on wrongly assigned haplotypes.

Since technological developments in genetics evolve rapidly, new kinds of genetic variants are currently fostered or will be available for epidemiological research in near future. For example, repeated sequences of base pairs can be counted, lining up successively in the genome. These variations are called *Copy Number Variations* (*CNV*) and are promising candidates for epidemiological studies. For such new measurement technologies, the quantification of measurement error is an important issue, which has to be taken into account in future studies. The perspective to use whole genome sequencing in epidemiological studies, which has been performed so far only on very limited samples worldwide, puts the measurement error topic on a very large scale.

6. Summary

This work focused on misclassification of genetic variants and its impact on genetic association studies. The initial question was, if and in which amount non-replication and inconsistency in these studies can be explained by errors in genotypes and reconstruction of haplotypes.

The amount and structure of genotyping errors were estimated via maximum-likelihood method based on double genotype measurements. These measurements were derived within routine quality control from genetic epidemiological association studies. Thereby it was possible to yield realistic error size estimates, as they can be expected in association analyses.

Genotyping error per SNP in a high quality laboratory using an established genotyping method has been found to be small (<0.1%). The data suggested the allelic drop out model to be appropriate and to some extent also the symmetric model. The bias due to genotype misclassification, as shown in a re-analysis of the association of SNPs of the *APM1* gene on plasma adiponectin concentrations, was found to be negligible. In other settings, e.g. relaxed quality control, genotype error might be higher. Then, a higher bias is expected, which was shown to be efficiently corrected with the MC-SIMEX method, a statistical method to correct for misclassification within a generalized linear model.

Regarding the uncertainties in haplotypes induced by statistical reconstruction from genotypes, a classification of the various haplotype error measures was provided, introducing sensitivity and specificity into the context of haplotypes. Results from simulations and analytical derivations emphasized the dependence of the haplotype reconstruction error on the specific situation, particularly on minor allele frequency, correlation between SNPs, number of loci and ambiguity fraction. Generally, the sensitivity was greatly reduced for some rare haplotypes, posing a potential problem of rare haplotypes in association studies.

Extension to a full 3x3 misclassification matrix, which has not been performed before in other methodological studies, allowed the inclusion of genotype errors. It could be shown that errors in genotypes add substantially to the pure reconstruction error.

The impact of haplotype misclassification, induced by a combination of genotype error and haplotype reconstruction error, on haplotype association analyses was evaluated in simulations as well as in a re-analysis of haplotypes on the *APM1* gene. In the case of a high genotype error per allele (1% or more), a rather high bias on haplotype association estimates was observed, which could be corrected using the MC-SIMEX method. The MC-SIMEX was

presented as an efficient method to calculate error-corrected association estimates in haplotype association studies.

Altogether, assuming good quality standards in the laboratory and thus a small genotype error rate (<0.5%) as it was estimated in this investigation, the impact on haplotype association estimates was rather moderate to negligable. Moreover, calculation of misclassification matrices for specific haplotypes additionally assures the correctness of haplotype assignments and simplifies the interpretation of association estimates.

These findings argue that non-replication of genetic association studies are only to a minor extent due to errors in genetic variants, if the genotyping process is performed in experienced laboratories using established methods with sufficient quality control.

The multiple testing problem is likely to play the biggest role in the non-replication problem of genetic epidemiological studies.

7. Zusammenfassung

Diese Arbeit befasst sich mit Fehlklassifikationen in genetischen Varianten und deren Einfluss auf genetische Assoziationsstudien. Ausgangspunkt war die Frage, in welchem Umfang nicht replizierbare und inkonstistente Ergebnisse in diesen Studien durch Fehler bei der Genotypisierung und der Berechnung von Haplotypen erklärt werden können.

Aus Doppeltmessungen wurde mit Hilfe eines Maximum-Likelihood-Verfahrens die Größe und Struktur des Genotypfehlers geschätzt. Diese Messungen standen aus Qualitätskontrollmaßnahmen laufender genetisch epidemiologischer Studien zur Verfügung. Dadurch konnte eine realistische Fehlerschätzung erreicht werden, wie sie in Assoziationsanalysen zu erwarten ist.

Der Genotypisierungsfehler pro SNP wurde bei einer etablierten Genotypisierungsmethode mit hohen Qualitätsstandards als sehr klein (<0.1%) geschätzt. Das geschätzte Fehlermodell entsprach dem allelischen Ausfallsmodell und wurde auch sehr gut durch das symmetrische Fehlermodell abgebildet. Bei einer Reanalyse der Assoziation von SNPs des *APM1*-Gens auf Adiponektin-Plasmawerte, zeigte sich nur eine minimale Verzerrung aufgrund der geschätzten Genotypfehlklassifikation, die als vernachlässigbar angesehen werden kann. Unter gelockerten Qualitätsstandards könnte der Genotypisierungsfehler allerdings höher ausfallen. In solchen Situationen wird eine stärkere Verzerrung erwartet. Es konnte allerdings gezeigt werden, das diese unter Verwendung der MC-SIMEX Methode, einer statistischen Methode zur Korrektur von Fehlklassifikationen, in effizienter Weise korrigiert werden kann.

Bezüglich der Unsicherheiten in Haplotypen durch die statistische Rekonstruktion wurden verschiedene Fehlermasse untersucht und klassifiziert, u.a. Sensitivität und Spezifität, die in diesem Kontext erstmalig verwendet wurden. Basierend auf Simulationen und analytischen Herleitungen wurden Abhängigkeiten des Haplotypfehlers von der Allelfrequenz und Anzahl und Korrelation der SNPs gefunden. Im Allgemeinen hat sich gezeigt, dass die Sensitivität und damit die Rekonstruktionsgenauigkeit v.a. für seltene Haplotypen eher klein war.

In bisherigen methodischen Analysen wurde die Möglichkeit der Genotypisierungsfehler bei der Haplotyprekonstruktion nicht miteinbezogen. In dieser Arbeit wurde die Fehlklassifikationsmatrix der Haplotypen dahingehend ausgeweitet. Es konnte gezeigt werden, dass Fehler in Genotypen substantiell zum Haplotypfehler beitragen.

Der Einfluß dieser Haplotypfehlklassifikation aus der Kombination von Genotypfehler und Rekonstruktionsfehler auf Assoziationsanalysen wurde in Simulationen und in einer

Reanalyse der *APM1*-Haplotypen untersucht. Im Falle eines hohen Genotypisierungsfehlers pro Allel (1% und größer) wurde eine substantielle Verzerrung der Schätzer für Haplotypen beobachtet. Mit der MC-SIMEX Methode kann allerdings für diese Haplotypfehlklassifikation, wie auch schon bei Genotypen allein, korrigiert werden. Diese Methode wurde in dieser Arbeit in den Kontext der genetischen Assoziationsanalyse v.a. bezüglich Haplotypen eingeführt und dafür evaluiert. Insgesamt ergibt sich, dass bei guten Qualitätsstandards des Labors und damit einem geringen Genotypisierungsfehler (<0.5%) eher moderate bis vernachlässigbare Verzerrungen in Haplotypassoziationsanalysen zu erwarten sind. Die Berechnung von Haplotypfehlklassifikationsmatrizen kann darüber hinaus zur Aussagekraft und Interpretation von Haplotypen und deren Assoziationssschätzer beitragen.

Diese Ergebnisse sprechen dafür, dass die vielen nicht-reproduzierbaren Ergebnisse genetisch epidemiologischer Studien eher auf andere Ursachen als auf Fehler in den genetischen Varianten zurückzuführen sind.

Appendix

A1 R-Function Sensitivity

```
################  Sensitivity  ###########################################################################
#
#		Function "Sensitivity" calculates Sensitivity, Specificity and Efficiency for each haplotype
#
#
#		Input parameter:
#
#			haplos:	Vector of haplotypes in 1/2-Coding, if code="1/2" (default):
#					e.g. 2-locus haplotype: haplos=c("11","12","21","22")
#				or, alternatively:
#				Matrix of haplotypes in 0/1-Coding, if code="0/1":
#					column 1: allele of locus 1, column 2: allele of locus 2 etc
#					row 1: haplotype 1, row 2: haploptype 2 etc.
#					e.g. 2- locus haplotype: haplos=matrix(c(0,0,1,1,0,1,0,1),ncol=2)
#
#			freqs:	vector of haplotype frequencies
#			code:	"1/2"(default), if input of haplotypes is a vector in "1/2"-Coding (see haplos)
#				"0/1", if input of haplotypes is a matrix in "0/1"-Coding (see haplos)
#
#		Output:
#
#			List of length=2
#				1. component:
#					Dataframe containing haplotypes (col 1), haplotype frequencies (col 2), sensitivity (col 3),
#					specificity (col 4) and efficiancy (col 5)
#
#				2. component:
#					Dataframe containing haplotypes (col 1) and probabilities of the misclassificatin matrix:
#					p11: Prob. that haplotype is truely present and assigned after estimation
#					p12: Prob. that haplotype is truely present but not assigned after estimation
#					p21: Prob. that haplotype is truely not present but after estimation
#					p22: Prob. that haplotype is truely not present and not assigned after estimation
#					p.1: Prob. that haplotype is assigned after estimation
#					p.2: Prob. that haplotype is not assigned after estimation
#					p1.: Prob. that haplotype is truely present
#					p2.: Prob. that haplotype is truely not present
#
#		Examples:
#
#			Example 1 (code="0/1"):
#
#			haplotypes <- matrix(c(0,0,1,1,0,1,0,1),ncol=2)
#			frequencies <- c(0.4,0.3,0.2,0.1)
#			Sensitivity(haplotypes,frequencies,"0/1")
#
#			Example 2 (code="1/2"):
#
#			haplotypes <- c("11","12","21","22")
#			frequencies <- c(0.4,0.3,0.2,0.1)
#			Sensitivity(haplotypes,frequencies)
#
#########################################################################################################################


Sensitivity <- function(haplos,freqs,code="1/2")
{
	freqs <- freqs / sum(freqs)		#Standardization of sum of haplotype frequencies to 1
	Leng <- length(freqs)

	if(code=="1/2")
	{
		haplo2 <- haplos
		Loci <- nchar(haplos[1])
		haplotypes <- matrix(rep(NA,Leng*Loci),ncol=Loci)
		for(i in 1:Loci)
		{
			haplotypes[,i] <- substring(haplos,i,i)
		}
		haplos <- matrix(as.integer(haplotypes),ncol=Loci) - 1
```

```
}
else
	haplo2 <- apply(haplos,1,paste,collapse="")

Loci <- length(haplos[1,])
er.h <- rep(NA,Leng)

haplos <- t(haplos)
genos <- matrix(rep(NA,Leng^2),ncol=Leng)
mono.freqs <- matrix(rep(NA,Leng^2),ncol=Leng)

for(i in 1:Leng)
{
		#genos: matrix with genotypes for each diplotype
		#e.g. "02" for diplotype 01/01 (0 minor alleles at locus 1, 2 minor alleles at locus 2)
	genos[i,] <- apply(haplos[,i] + haplos, 2, paste, collapse="")

		#mono.freqs: frequencies of (sorted) diplotypes assuming HWE
	mono.freqs[i,] <- freqs[i] * freqs
}

genotypes <- sort(unique(genos))		#all present genotypes
gen.freqs <- rep(NA,length(genotypes))

	#gen.freqs: sum of all diplotype-frequencies with the same genotype
for(i in 1:length(genotypes))
	gen.freqs[i] <- sum((genos==genotypes[i]) * mono.freqs)
	#genos: matrix, holding genotype frequencies

genos2 <- as.factor(genos)
genos <- as.factor(genos)
Levels <- matrix(as.integer(genos),ncol=Leng)
levels(genos) <- gen.freqs
genos <- matrix(as.numeric(as.character(genos)),ncol=Leng)

	#mono.freqs: frequencies of sorted diplotypes (e.g.: P(h1/h2)=p and P(h2/h1)=p)
	#diplo.freqs: frequencies of diplotypes without consideration of order (e.g.: P(h1 and h2)=2*p)

right <- matrix(rep(0,Leng^2),ncol=Leng)

for(i in levels(genos2))
	right <- right + matrix(((genos2 == i) * mono.freqs) == max((genos2 == i) * mono.freqs),ncol=Leng)

diplo.freqs <- mono.freqs + (1 - diag(rep(1,Leng))) * mono.freqs

	#Calculation of misclassification and marginal probabilities

p1. <- apply(diplo.freqs, 1, sum)
p11 <- apply(diplo.freqs * (right), 1, sum)
p12 <- p1. - p11

p.1 <- rep(0,Leng)

for(j in 1:Leng)
{
	for(i in 1:Leng)
	{
		p.1[j] <- p.1[j] + (right[i,j]) * sum(((Levels == Levels[i,j]) * mono.freqs))
	}
}

p.2 <- 1 - p.1
p21 <- p.1 - p11
p22 <- p.2 - p12
p2. <- 1 - p1.

	#Calculation of Sensitivity, Specificity, Efficiency

sensitivity <- p11 / p1.
specificity <- p22 / p2.
efficiency <- p11 + p22

list(data.frame(Haplotype=haplo2,Frequency=freqs,Sensitivity=sensitivity,Specificity=specificity,Efficiency=efficiency),
data.frame(Haplotype=haplo2,p11=p11,p12=p12,p21=p21,p22=p22,p.1=p.1,p.2=p.2,p1.=p1.,p2.=p2.))
}
```

```
#                                                                                                                    END
#####################################################################################################################
```

A2 R-Function Starplot

```
######## Starplot ###################################################################################################
#
#       Function "Starplot" draws Sensitivity, Specificity or R-Square-values for each haplotype
#
#
#       required arguments:
#               values:             vector of parameter values to be plotted for each haplotype represented by the length of the line
#                                   in the circle
#                                   (e.g. the sensitivity or specificity given by the function "sensitivity", or
#                                   the R-square that can be calculated by the program tagSNPs by Dan Stram)
#
#       optional arguments:
#               labels:             vector of labels to be printed for each haplotype
#               freqs:              vector of frequencies of each haplotype;
#                                   if frequencies are specified, they will be printed in the plot
#                                   if frequencies are not specified, haplotypes are plotted in the given order
#               orderplot:  if orderplot=T (default), haplotypes are plotted in the order of their frequencies
#                                   if orderplot=F, haplotypes are plotted in the given order
#               circles:    vector of radii of circles to be added to the plot. By default, one circle with
#                                   radius=1 is printed.
#               circ.labels:if circ.labels=T: the circle radii are specified
#                                   if circ.labels=F (default): the circle radii are not specified
#               label.dist:  Distance of the labels to the center of the plot. By default, this is 1.15
#               label.size:  Size of the labels (default=1)
#
#               Graphical parameters may also be supplied to this function (see S-Plus-help for 'par'), e.g. main="" as main title
#
#
#       Examples:
#
#               Example 1:
#
#               starplot(c(0.9,0.8,0.6,0.5), c("AA","AB","BA","BB"), freqs=c(0.5,0.3,0.15,0.05))
#
#               Example 2:
#
#               haplotypes <- c("000", "001", "010", "011", "100", "101", "110", "111")
#               frequencies <- c(0.48000,0.05408,0.24805,0.00000,0.17480,0.01403,0.02448,0.00000)
#               sens <- c(0.9440987,0.8767955,0.9381565,NA,0.9069180,0.5810413,0.5647062,NA)
#               starplot(sens, haplotypes, frequencies)
#
#               Example 3 (using the output of the function "sensitivity" to plot the sensitivity and specificity):
#
#               haplotypes <- matrix(c(0,0,1,1,0,1,0,1),ncol=2)
#               frequencies <- c(0.4,0.3,0.2,0.1)
#               example.plot <- Sensitivity(haplotypes,frequencies,"0/1")
#               par(mfrow=c(1,2))
#               starplot(example.plot[[1]]$Sensitivity, example.plot[[1]]$Haplotype, example.plot[[1]]$Frequency,
#               main="Sensitivity", label.dist=1.17, label.size=0.7)
#               starplot(example.plot[[1]]$Specificity, example.plot[[1]]$Haplotype, example.plot[[1]]$Frequency,
#               main="Specificity", label.dist=1.17, label.size=0.7)
#
#####################################################################################################################


starplot <- function(values,labels=0,freqs=0,orderplot=T,circles=1,circ.labels=F,label.dist=1.15,label.size=1,...)
{
        plotfreqs <- T
        if(length(labels)==1) labels <- rep("",length(values))
        if(length(freqs)==1)
                {
                freqs <- rep("",length(values))
                orderplot <- F
                plotfreqs <- F
                }

        if(orderplot==T)
```

```
		{
		variable <- list(values,as.character(labels),freqs)
		orderlist <- order(-variable[[3]])
		}
	else
		{
		variable <- list(values,as.character(labels),freqs)
		orderlist <- 1:4
		}

	circles <- rev(sort(circles))
	par(pty="s")
	plot(complex(modulus=1.2,argument=seq(-pi, pi, length=1000)),type="n",axes=F,xlab="",ylab="",...)
	lines(complex(modulus=circles[1],argument=seq(-pi, pi, length=1000)))
	if(circ.labels) text(circles[1]+0.03*circles[1],0.04*circles[1],circles[1])

	if(length(circles) > 1)
	{
	for(i in 2:length(circles))
		{
		lines(complex(modulus=circles[i],argument=seq(-pi, pi, length=1000)))
		if(circ.labels) text(circles[i]+0.03*circles[i],0.04*circles[i],circles[i])
		}
	}

	numvariables <- length(values)
	argum <- (0:(numvariables-1)) / numvariables * 2 * pi

	sine <- sin(argum)
	cosine <- cos(argum)

	sine.lab <- sine * label.dist
	cosine.lab <- cosine * label.dist
	sine <- sine*variable[[1]][orderlist]
	cosine <- cosine*variable[[1]][orderlist]

	if (plotfreqs==T)
	{
	for(i in 1:length(labels))
		{
		lines(c(0,sine[i]),c(0,cosine[i]))
		text(sine.lab[i],cosine.lab[i],paste(variable[[2]][orderlist][i],"\n
",round(100*variable[[3]][orderlist][i],1),"%",sep=""),cex=label.size)
		}
	}
	else
	{
	for(i in 1:length(labels))
		{
		lines(c(0,sine[i]),c(0,cosine[i]))
		text(sine.lab[i],cosine.lab[i],variable[[2]][orderlist][i],cex=label.size)
		}
	}
}

#                                                                                                   END
###################################################################################################################
```

A3 Frequency, sensitivity, and specificity for *APM1* haplotypes

All haplotypes with frequency >1% assuming 0, 0.25, 0.5 and 1% genotype error per allele

		Genotyping Error							
		0%		**0.25%**		**0.5%**		**1%**	
Haplotype	freq	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
H22	0.124	0.9743	0.9644	0.9065	0.9593	0.8385	0.956	0.7271	0.9497
H16	0.100	0.9585	0.9726	0.8899	0.9673	0.8225	0.9642	0.7084	0.9589
H26	0.096	0.9957	0.9989	0.9235	0.9986	0.8533	0.9983	0.7408	0.998
H15	0.093	0.8565	0.9952	0.7926	0.9909	0.7339	0.9872	0.6333	0.9811
H11	0.066	0.9337	0.986	0.8592	0.9854	0.7955	0.9841	0.6824	0.9832
H9	0.066	0.9965	0.9888	0.9203	0.9885	0.8549	0.9877	0.7337	0.9872
H8	0.054	0.9956	0.997	0.9275	0.996	0.8553	0.995	0.7364	0.9934
H2	0.053	0.9932	0.9947	0.9252	0.9947	0.8587	0.9944	0.7294	0.9941
H24	0.044	0.983	0.9945	0.9117	0.9947	0.8441	0.9945	0.7279	0.9945
H23	0.039	0.7855	0.9929	0.7152	0.9929	0.6548	0.9928	0.5574	0.9929
H19	0.031	0.8483	0.9894	0.7898	0.9884	0.726	0.9874	0.629	0.9862
H21	0.025	0.9874	1	0.918	0.9987	0.8514	0.9973	0.7352	0.9947
H12	0.023	0.8986	0.9947	0.8288	0.9941	0.7575	0.9936	0.6544	0.9929
H20	0.019	0.8628	0.9952	0.7927	0.9954	0.7315	0.9951	0.6267	0.9951
H4	0.019	0.9816	1	0.9086	0.9987	0.8441	0.9976	0.7299	0.9955
H1	0.016	0.8915	0.9993	0.839	0.999	0.7546	0.9992	0.6476	0.9989
H17	0.011	0.9847	1	0.9105	0.9995	0.8504	0.9991	0.7209	0.9983
H6	0.010	0.9811	0.9998	0.9155	0.9994	0.8398	0.9992	0.6946	0.9988

A4 Misclassification matrices for *APM1* haplotypes

All haplotypes with frequency >1% assuming 0, 0.25, 0.5 and 1% genotype error per allele

		Genotyping Error															
Haplotype	freq	**0%**				**0.25%**				**0.5%**				**1%**			
H22	0.124		True Haplo H				True Haplo H				True Haplo H				True Haplo H		
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9644	0.0275	0	0	0.9593	0.0565	0	0	0.956	0.089	0.0025	0	0.9497	0.1394	0.0036
		C 1	0.0356	0.9725	0	C 1	0.0407	0.943	0.0613	C 1	0.044	0.9105	0.1275	C 1	0.0503	0.8599	0.2266
		2	0	0	1	2	0	0.0004	0.9387	2	0	0.0004	0.87	2	0	0.0007	0.7698
H16	0.100		True Haplo H				True Haplo H				True Haplo H				True Haplo H		
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9726	0.0436	0	0	0.9673	0.0773	0	0	0.9642	0.1141	0	0	0.9589	0.1754	0.0058
		C 1	0.0274	0.9564	0	C 1	0.0327	0.9225	0.0838	C 1	0.0358	0.8857	0.1438	C 1	0.0411	0.824	0.2406
		2	0	0	1	2	0	0.0002	0.9162	2	0	0.0002	0.8562	2	0	0.0006	0.7535
H26	0.096		True Haplo H				True Haplo H				True Haplo H				True Haplo H		
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9989	0.0045	0	0	0.9986	0.0299	0	0	0.9983	0.0597	0	0	0.998	0.1165	0
		C 1	0.0011	0.9955	0	C 1	0.0014	0.9701	0.0656	C 1	0.0017	0.9403	0.147	C 1	0.002	0.8834	0.2714
		2	0	0	1	2	0	0	0.9344	2	0	0	0.853	2	0	0.0001	0.7286
H15	0.093		True Haplo H				True Haplo H				True Haplo H				True Haplo H		
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9989	0.1508	0	0	0.9909	0.1826	0	0	0.9872	0.2148	0	0	0.9811	0.2696	0.0015
		C 1	0.0048	0.8492	0	C 1	0.0091	0.817	0.0874	C 1	0.0128	0.7846	0.1536	C 1	0.0189	0.7297	0.3
		2	0	0	1	2	0	0.0005	0.9126	2	0	0.0006	0.8464	2	0	0.0007	0.6985

H11	0.066		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.986	0.0688	0		0	0.9854	0.1213	0		0	0.9841	0.1606	0		0	0.9832	0.2268	0.0113
	C	1	0.014	0.9312	0	C	1	0.0146	0.8779	0.0353	C	1	0.0159	0.8382	0.0958	C	1	0.0168	0.7716	0.1961
		2	0	0	1		2	0	0.0008	0.9647		2	0	0.0012	0.9042		2	0	0.0016	0.7926
H9	0.066		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9888	0.0036	0		0	0.9885	0.0469	0		0	0.9877	0.0801	0		0	0.9872	0.1445	0
	C	1	0.0112	0.9964	0	C	1	0.0115	0.9531	0.0682	C	1	0.0123	0.9199	0.1218	C	1	0.0128	0.8553	0.2744
		2	0	0	1		2	0	0	0.9318		2	0	0	0.8782		2	0	0.0001	0.7256
H8	0.054		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.997	0.0045	0		0	0.996	0.0512	0		0	0.995	0.0969	0.0072		0	0.9934	0.1644	0.0072
	C	1	0.003	0.9955	0	C	1	0.004	0.9488	0.0649	C	1	0.005	0.9029	0.1025	C	1	0.0066	0.8354	0.2301
		2	0	0	1		2	0	0	0.9351		2	0	0.0002	0.8902		2	0	0.0002	0.7627
H2	0.053		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9947	0.007	0		0	0.9947	0.0535	0		0	0.9944	0.0852	0		0	0.9941	0.1672	0
	C	1	0.0053	0.993	0	C	1	0.0053	0.9465	0.0278	C	1	0.0056	0.9148	0.0934	C	1	0.0059	0.8328	0.1833
		2	0	0	1		2	0	0	0.9722		2	0	0	0.9066		2	0	0	0.8167
H24	0.044		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9945	0.0176	0		0	0.9947	0.0656	0		0	0.9945	0.1057	0		0	0.9945	0.175	0
	C	1	0.0055	0.9824	0	C	1	0.0053	0.9344	0.0505	C	1	0.0055	0.8943	0.1124	C	1	0.0055	0.825	0.2223
		2	0	0	1		2	0	0	0.9495		2	0	0	0.8876		2	0	0	0.7777

H23	0.039		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9929	0.2183	0	0	0.9929	0.2704	0	0	0.9928	0.3136	0	0	0.9929	0.3786	0
		C 1	0.0071	0.7817	0	C 1	0.0071	0.7296	0.1343	C 1	0.0072	0.6864	0.2292	C 1	0.0071	0.6214	0.3704
		2	0	0	1	2	0	0	0.8657	2	0	0	0.7708	2	0	0	0.6296
H19	0.031		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9894	0.1545	0	0	0.9884	0.2008	0	0	0.9874	0.248	0	0	0.9862	0.3216	0.0098
		C 1	0.0106	0.8455	0	C 1	0.0116	0.7992	0.0794	C 1	0.0126	0.7513	0.2093	C 1	0.0138	0.6762	0.3672
		2	0	0	1	2	0	0	0.9206	2	0	0.0007	0.7907	2	0	0.0022	0.623
H21	0.025		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	1	0.0128	0	0	0.9987	0.0715	0	0	0.9973	0.1238	0	0	0.9947	0.2158	0
		C 1	0	0.9872	0	C 1	0.0013	0.9282	0.0962	C 1	0.0027	0.8759	0.1346	C 1	0.0053	0.783	0.2115
		2	0	0	1	2	0	0.0004	0.9038	2	0	0.0004	0.8654	2	0	0.001	0.7885
H12	0.023		True Haplo H				True Haplo H				True Haplo H				True Haplo H			
			0	1	2		0	1	2		0	1	2		0	1	2
		0	0.9947	0.103	0	0	0.9941	0.1571	0	0	0.9936	0.2096	0	0	0.9929	0.2902	0.0345
		C 1	0.0053	0.897	0	C 1	0.0059	0.8425	0.1034	C 1	0.0064	0.79	0.2586	C 1	0.0071	0.7091	0.3276
		2	0	0	1	2	0	0.0004	0.8966	2	0	0.0004	0.7414	2	0	0.0007	0.6379
H20	0.019																	

		True Haplo H			True Haplo H			True Haplo H			True Haplo H		
		0	1	2	0	1	2	0	1	2	0	1	2
	0	0.9952	0.1381	0	0	0.9954	0.1983	0	0	0.9951	0.2399	0	0	0.9951	0.3132	0
C	1	0.0048	0.8619	0	C 1	0.0046	0.8017	0.1026	C 1	0.0049	0.7601	0.1026	C 1	0.0049	0.6868	0.4872
	2	0	0	1	2	0	0	0.8974	2	0	0	0.8974	2	0		0.5128
H4	0.019	True Haplo H			True Haplo H			True Haplo H			True Haplo H		
		0	1	2	0	1	2	0	1	2	0	1	2
	0	1	0.0185	0	0	0.9987	0.0804	0	0	0.9976	0.1363	0	0	0.9955	0.2265	0
C	1	0	0.9815	0	C 1	0.0013	0.9196	0	C 1	0.0024	0.8637	0	C 1	0.0045	0.7735	0
	2	0	0	1	2	0	0	1	2	0	0	1	2	0	0	1
H1	0.016	True Haplo H			True Haplo H			True Haplo H			True Haplo H		
		0	1	2	0	1	2	0	1	2	0	1	2
	0	0.9993	0.1097	0	0	0.999	0.1525	0	0	0.9992	0.2162	0	0	0.9989	0.2948	0
C	1	0.0007	0.8903	0	C 1	0.001	0.8475	0.0294	C 1	0.0008	0.7838	0.0882	C 1	0.0011	0.7052	0.2941
	2	0	0	1	2	0	0	0.9706	2	0	0	0.9118	2	0	0	0.7059
H17	0.011	True Haplo H			True Haplo H			True Haplo H			True Haplo H		
		0	1	2	0	1	2	0	1	2	0	1	2
	0	1	0.0153	0	0	0.9995	0.0782	0	0	0.9991	0.1309	0	0	0.9983	0.2347	0
C	1	0	0.9847	0	C 1	0.0005	0.9218	0.5	C 1	0.0009	0.8691	0.5	C 1	0.0017	0.7653	0.5
	2	0	0	1	2	0	0	0.5	2	0	0	0.5	2	0	0	0.5
H6	0.010	True Haplo H			True Haplo H			True Haplo H			True Haplo H		
		0	1	2	0	1	2	0	1	2	0	1	2
	0	0.9998	0.019	0	0	0.9994	0.073	0	0	0.9992	0.1365	0	0	0.9988	0.2553	0
C	1	0.0002	0.981	0	C 1	0.0006	0.927	0	C 1	0.0008	0.8635	0	C 1	0.0012	0.7447	0.25
	2	0	0	1	2	0	0	1	2	0	0	1	2	0	0	0.75

A5 Related publications and description of own contribution

The content of this thesis has already given rise to three publication manuscripts:

- Lamina C, Bongardt F, Küchenhoff H, Heid IM. Haplotype Reconstruction Error as a Classical Misclassification Problem: Introducing Sensitivity and Specificity as Error Measures. PLoS ONE. 2008 Mar 26;3(3):e1853
- Heid IM, Lamina C, Küchenhoff H, Fischer G, Klopp N, Kolz M, Grallert H, Vollmert C, Wagner S, Huth C, Müller J, Hunt SC, Peters A, Paulweber B, Wichmann HE, Kronenberg F, Illig T. Estimating the SNP Genotype Misclassification from Routine Double Measurements in a large Epidemiological Sample. American Journal of Epidemiology, Accepted.
- Lamina C, Küchenhoff H, Chang-Claude J, Paulweber B, Wichmann HE, Illig T, Hoehe MR, Kronenberg F, Heid IM. Haplotype Misclassification from Genotype Error and Statistical Reconstruction and its impact on association estimates. Submitted

Several authors contributed to these manuscripts with Dr. Iris Heid being the first author of one manuscript. Therefore, the own and original contribution of the author of these manuscripts are explained in the following:

Claudia Lamina contributed to great extent to the scientific concept and manuscript writing to all three manuscripts. She planned and conducted most of the statistical analysis and drew conclusions.

The project on genotype and haplotype uncertainties was initiated by Prof. Dr. Dr. Wichmann, Dr. Iris Heid, Prof. Dr. Florian Kronenberg and Dr. Thomas Illig as a subproject within the SFB 386 (Sonderforschungsbereich 386: Statistical analysis of discrete structures) of the LMU Munich in cooperation with the Institute of Statistics, and the National Genome Network (NGFN).

Regarding the first manuscript, Friedhelm Bongardt provided a first version of the R-programs “Sensitivity” and “Starplot”, which were adapted and finalized by the author of this thesis. Further analytical solutions, literature search, figures and manuscript writing were performed by the author of this thesis.

Claudia Lamina contributed to the scientific concept for the second manuscript, supervised by Dr. Iris Heid. The formulas for maximum-likelihood estimation were designed by Dr. Iris Heid, PD Dr. Helmut Küchenhoff and Claudia Lamina in close cooperation. The author of this thesis implemented this estimation process in Mathematica in sole responsibility. All analysis on the genotyping error manuscript, descriptive statistics and reanalysis of the *APM1*-gene, was conducted by Claudia Lamina. Manuscript writing was carried together with

Dr. Iris Heid. The other authors of this manuscript, Guido Fischer, Norman Klopp , Melanie Kolz, Harald Grallert, Caren Vollmert, Stefanie Wagner, Cornelia Huth, Steve Hunt, Anette Peters and Bernhard Paulweber contributed to this investigation by providing the duplicate genotype data from the genotyping facility, from the data management side or as principal investigators of the corresponding studies.

Regarding the third manuscript, the extension of haplotype misclassification to include genotype error, Claudia Lamina performed all simulations and data analysis with main responsibility on literature search, figures and manuscript writing.

All coauthors of the publication manuscripts revised the manuscript drafts and some gave advice on language, comprehensibility and statistical analysis.

Although main parts of this thesis were based on the manuscripts mentioned, it is not a copy of these manuscripts. The majority of the introduction and many parts of the main sections and discussion are original to this thesis and has not been part of any of the publication manuscripts.

A6 List of Publications and Presentations

Original articles

Lamina C, Meisinger C, Heid IM, Rantner B, Döring A, Löwel H, Wichmann HE, Kronenberg F. Ankle-brachial index and peripheral arterial disease. Gesundheitswesen. 2005. 67 Suppl 1:S57-61.

Lamina C, Steffens M, Mueller J, Lohmussaar E, Meitinger T, Wichmann HE. Genetic diversity in German and European populations: looking for substructures and genetic patterns. Gesundheitswesen. 2005. 67 Suppl 1:S127-31.

Heid IM, **Lamina** C, Bongardt F, Fischer G, Klopp N, Huth C, Küchenhoff H, Kronenberg F, Wichmann HE, Illig T. How About the Uncertainty in the Haplotypes in the Population-Based KORA Studies. Gesundheitswesen. 2005. 67 Suppl 1: S132–136.

Schwaiger JP, **Lamina C**, Neyer U, König P, Kathrein H, Sturm W, Lhotta K, Gröchenig E, Dieplinger H, Kronenberg F. Carotid plaques and their predictive value for cardiovascular disease and all-cause mortality in hemodialysis patients considering renal transplantation: a decade follow-up. American Journal of Kidney Disease. 2006 May;47(5):888-97.

Lamina C, Meisinger C, Heid IM, Löwel H, Rantner B, Koenig W, Kronenberg F; Kora Study Group. Association of ankle-brachial index and plaques in the carotid and femoral arteries with cardiovascular events and total mortality in a population-based study with 13 years of follow-up. European Heart Journal. 2006 Nov;27(21):2580-7.

Steffens M, **Lamina C**, Illig T, Bettecken T, Vogler R, Entz P, Suk EK, Toliat MR, Klopp N, Caliebe A, König IR, Köhler K, Ludemann J, Diaz Lacava A, Fimmers R, Lichtner P, Ziegler A, Wolf A, Krawczak M, Nürnberg P, Hampe J, Schreiber S, Meitinger T, Wichmann HE, Roeder K, Wienker TF, Baur MP. SNP-based analysis of genetic substructure in the German population. Human Heredity. 2006;62(1):20-9.

Vollmert C, Hahn S, **Lamina** C, Huth C, Kolz M, Schopfer-Wendels A, Mann K, Bongardt F, Mueller JC, Kronenberg F, Wichmann HE, Herder C, Holle R, Lowel H, Illig T, Janssen OE; the KORA group. Calpain-10 variants and haplotypes are associated with polycystic ovary syndrome in Caucasians. American Journal of Physiology, Endocrinology and Metabolism. 2007. 292(3): E836–844.

Lamina C, Bongardt F, Küchenhoff H, Heid IM. Haplotype Reconstruction Error as a Classical Misclassification Problem: Introducing Sensitivity and Specificity as Error Measures. PLoS ONE. 2008 Mar 26;3(3):e1853

Müller M, Döring A, Küchenhoff H, **Lamina C**, Malzahn D, Bickeböller H, Vollmert C, Klopp N, Meisinger C, Heinrich J, Kronenberg F, Wichmann HE, Heid IM: Quantifying the contribution of genetic variants for survival phenotypes. Genet Epidemiol. 2008 Sep;32(6):574-85.

Loos RJF, Lindgren CM, Li S, …, Heid IM, KORA*,…, McCarthy M, Wareham NJ, Barroso I. Association studies involving over 90000 samples demonstrate that common variants near to MC4R influence fat mass, weight and risk of obesity. Nat Genet. 2008 Jun;40(6):768-75
* Additional consortium authors: KORA: **Lamina C**, Gieger C, Illig T, Meitinger T, Wichmann HE

Linsel-Nitschke P, Götz A, Erdmann J, Braenne I, Lieb W, Hall AS, Hengstenberg C, Stark K, Schreiber S, El Mokhtari NE, Schaefer A, Schrezenmeier J, Rubin D, Hinney A, Reinehr T, Roth C, Schreiner F, Ortlepp J, Hanrath P, Braund P, Mangino M, **Lamina C**, Heid IM, Doering A, Gieger C, Peters A, Meitinger T, Wichmann HE, König IR, Ziegler A, Kronenberg F, Samani NJ, Schunkert H. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease--a Mendelian Randomisation study. PLoS ONE 2008 Aug 20;3(8):e2986

Heid IM, **Lamina C**, Küchenhoff H, Fischer G, Klopp N, Kolz M, Grallert H, Vollmert C, Wagner S, Huth C, Müller J, Hunt SC, Peters A, Paulweber B, Wichmann HE, Kronenberg F, Illig T. Estimating the SNP Genotype Misclassification from Routine Double Measurements in a large Epidemiological Sample. Am J Epidemiol. 2008 Oct 15;168(8):878-89

Heid IM, Boes E, Müller M, Kollerits B, **Lamina C** , Coassin S, Gieger C, Döring A, Klopp N, Frikke-Schmidt R, Tybjærg-Hansen A, Brandstätter A, Luchner A, Meitinger T, Wichmann HE, Kronenberg F: A Genome-Wide Association Analysis of HDL-Cholesterol in the Population-Based KORA Study Sheds New Light on Intergenic Regions. Circ Cardiovasc Genet. 2008;1:10-20

Willer C, Speliotes E, Loos R, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, **Lamina C**,…,Wichmann HE, McCarthy MI, Boehnke M, Barroso I, Gonçalo A, Hirschhorn JN. Six New Loci Associated with Body Mass Index Highlight a Neuronal Influence on Body Weight Regulation. Nat Genet. Accepted.

Laumen H, Heid IM, Hess J, Herder C, **Lamina C**, Rathmann W, Saningong AD, Sedlmaier EM, Klopp N, Thorand B, Wichmann HE, Illig T, Hauner H. Functional characterization of promoter variants of the adiponectin gene complemented by epidemiological data. Submitted

Lamina C, Küchenhoff H, Chang-Claude J, Paulweber B, Wichmann HE, Illig T, Hoehe MR, Kronenberg F, Heid IM. Haplotype Misclassification from Genotype Error and Statistical Reconstruction and its impact on association estimates. Submitted

Marquard V, Beckmann L, Heid IM, **Lamina C**, Chang-Claude J: Impact of genotyping errors on the type I error and the power of haplotype-based association methods. Submitted

Presentations (Orally and Poster)

Lamina C. Meßfehler in der Genotyp-und Haplotyp-Assoziationsanalyse. Oral presentation at the Seminar on "Measurement Error Measures", Department of Statistics, LMU München, July 19, 2004

Lamina C, Kronenberg F, Heid IM, Meisinger C, Löwel H, Wichmann HE, Koenig W. Assoziation zwischen Ankle-Brachial-Index und atherosklerotischen Plaques mit Koronarereignissen in der MONICA Augsburg Studie (1989/90). Oral Presentation at the 49. annuel meeting of the Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds), Innsbruck, Austria, September 14, 2004

Lamina C, Kronenberg F, Heid IM, Meisinger C, Löwel H, Wichmann HE, Koenig W. Assoziation zwischen Ankle-Brachial-Index und Koronarereignissen in der MONICA Augsburg Studie. Oral Presentation at the Workshop „Extended survival Methods" of the Working group „Statistical methods in Epidemiology" of the German Association of Epidemiology (DAE), Halle, November 19, 2004

Lamina C, Küchenhoff H, Bongardt F, Paulweber B, Kronenberg F, Wichmann HE, Heid IM On the haplotype uncertainty from genotyping and reconstruction error and its impact on association analysis. Oral presentation at the Annual Conference of the Genetic Epidemiological Society (IGES), Park City, Utah, USA, October 24, 2005

Lamina C, Hahn S, Vollmert C, Kolz M, Schöpfer-Wendels A, Mann K, Bongardt F, Müller JC, Kronenberg F, Wichmann HE, Herder C, Holle R, König W, H. Löwel1, T. Illig T, Janssen OE and the KORA group. Calpain 10 and Interleukin 6 variants are associated with polycystic ovary syndrome. Poster presentation at the annual conference of the American Society of Human Genetics (ASHG), Salt Lake City, Utah, USA, October 25-29, 2005

Lamina C, Bongardt F, Heid IM, Wichmann HE. Haplotypes for association analysis. Oral presentation at the meeting of the SMP-DNA subgroup of the German National Genome Network (NGFN), Berlin, November 11, 2005

Lamina C. Genetische Assoziationsanalysen unter Berücksichtigung von Meßfehlern in Genotypen und Haplotypen. Oral presentation at the Seminar on "Measurement Error Measures", Department of Statistics, LMU München, December 16, 2005

Lamina C. Genetische Assoziationsanalysen unter Berücksichtigung von Meßfehlern in Genotypen und Haplotypen. Oral presentation at the workshop of the SFB 386 „Statistical Analysis of Discrete Structures", Höhenried, July 4, 2006

Lamina C, Küchenhoff H, Paulweber B, Bickeböller H, Illig T, Kronenberg F, Wichmann HE, Heid IM On the haplotype uncertainty from genotyping and reconstruction error and its impact on association analysis. Poster presentation at the Annual Meeting of the German National Genome Network (NGFN), Heidelberg, November 25–26, 2006

Lamina C, Heid IM, Küchenhoff H, Paulweber B, Bickeböller H, Illig T, Kronenberg F, Wichmann. The impact of haplotype uncertainty resulting from genotype error and statistical haplotype reconstruction on association analysis. Oral Presentation at the 2nd

Annual Meeting of the German Epidemiological Association, Augsburg, September 18, 2007

Lamina C, Abecasis G, Wichmann HE, Kronenberg F, Heid IM. Do we gain from imputing ungenotyped SNPs in genomewide association studies?. Submitted to the Annual Biometrical Colloquium "Statistics and Life Sciences", München, March 10-13, 2008

Lamina C. Schätzung des Genotyp- und Haplotypfehlers und deren Einfluß auf die Haplotypassoziationsanalyse. Invited talk at the "Mathematisches Kolloquium", University of Bremen, April 8, 2008

Lamina C, Heid IM, Boes E, Müller M, Kollerits B, Coassin S, Gieger C, Döring A, Klopp N, Frikke-Schmidt R, Tybjærg-Hansen A, Brandstätter A, Luchner A, Meitinger T, Wichmann HE, Kronenberg F: A Genome-Wide Association Analysis of HDL-Cholesterol in the Population-Based KORA Study Sheds New Light on Intergenic Regions. Poster presentation at the 3rd ESF Functional Genomics Conference, Innsbruck, Austria, October 1-4, 2008

A7 References

Adkins RM. 2004. Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. BMC Genet 5:22

Akey J, Jin L, Xiong M. 2001a. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291-300.

Akey JM, Zhang K, Xiong M, Doris P, Jin L. 2001b. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet 68:1447-1456.

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES. 2000. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 26:76-80.

Becker T, Valentonyte R, Croucher PJ, Strauch K, Schreiber S, Hampe J, Knapp M. 2006. Identification of probable genotyping errors by consideration of haplotypes. Eur J Hum Genet 14:450-458.

Bickeböller H, Fischer C. 2007. Einführung in die Genetische Epidemiologie.

Bross I. 1954. Misclassification in 2x2 tables. Biometrics 10:478-486.

Carroll RJ, Ruppert D, Stefanski LA. 2006. Measurement Error in Nonlinear Models.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS. 2007. Replicating genotype-phenotype associations. Nature 447:655-660.

Clark AG. 2004. The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321-333.

Cook JR, Stefanski LA. 1994. Simulation-Extrapolation Estimation in Parametric Measurement error Models. Journal of American Statistical Association 89:1314-1328.

Cordell HJ, Clayton DG. 2005. Genetic association studies. Lancet 366:1121-1131.

Cox DG, Kraft P. 2006. Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. Hum Hered 61:10-14.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. Nat Genet 29:229-232.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311-322.

Dosemeci M, Wacholder S, Lubin JH. 1990. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 132:746-748.

Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP. 2004. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. Am J Gastroenterol 99:2393-2404.

Epstein MP, Satten GA. 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.

Fallin D, Schork NJ. 2000. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947-959.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Gordon D, Finch SJ, Nothnagel M, Ott J. 2002. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered 54:22-33.

Gordon D, Haynes C, Yang Y, Kramer PL, Finch SJ. 2007. Linear trend tests for case-control genetic association that incorporate random phenotype and genotype misclassification error. Genet Epidemiol 31:853-870.

Gordon D, Ott J. 2001. Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac Symp Biocomput18-29.

Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM, Haroutunian V. 2004. Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. Stat Appl Genet Mol Biol 3:Article26

Govindarajulu US, Spiegelman D, Miller KL, Kraft P. 2006. Quantifying bias due to allele misclassification in case-control studies of haplotypes. Genet Epidemiol 30:590-601.

Gustafson P. 2003. Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.

Hao K, Wang X. 2004. Incorporating individual error rate into association test of unmatched case-control design. Hum Hered 58:154-163.

Heid IM, Wagner SA, Gohlke H, Iglseder B, Mueller JC, 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.

Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, Masson G, Gudbjartsson DF, Magnusson KP, Andersen K, Levey AI, Backman VM, Matthiasdottir S, Jonsdottir T, Palsson S, Einarsdottir H, Gunnarsdottir S, Gylfason A, Vaccarino V, Hooper WC, Reilly MP, Granger CB, Austin H, Rader DJ, Shah SH, Quyyumi AA, Gulcher JR, Thorgeirsson G, Thorsteinsdottir U, Kong A, Stefansson K. 2007. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316:1491-1493.

Helgason A, Palsson S, Thorleifsson G, Grant SF, Emilsson V, Gunnarsdottir S, Adeyemo A, Chen Y, Chen G, Reynisdottir I, Benediktsson R, Hinney A, Hansen T, Andersen G, Borch-Johnsen K, Jorgensen T, Schafer H, Faruque M, Doumatey A, Zhou J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Sigurdsson G, Hebebrand J, Pedersen O, Thorsteinsdottir U, Gulcher JR, Kong A, Rotimi C, Stefansson K. 2007. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat Genet 39:218-225.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. Genet Med 4:45-61.

Illig T, Bongardt F, Schopfer A, Holle R, Muller S, Rathmann W, Koenig W, Meisinger C, Wichmann HE, Kolb H. 2003. The endotoxin receptor TLR4 polymorphism is not associated with diabetes or components of the metabolic syndrome. Diabetes 52:2861-2864.

Illig T, Bongardt F, Schopfer A, Muller-Scholze S, Rathmann W, Koenig W, Thorand B, Vollmert C, Holle R, Kolb H, Herder C. 2004. Significant association of the interleukin-6 gene polymorphisms C-174G and A-598G with type 2 diabetes. J Clin Endocrinol Metab 89:5053-5058.

Ioannidis JP. 2007. Non-replication and inconsistency in the genome-wide association setting. Hum Hered 64:203-213.

Ioannidis JP, Patsopoulos NA, Evangelou E. 2007. Heterogeneity in meta-analyses of genome-wide association investigations. PLoS ONE 2:e841

Kang H, Qin ZS, Niu T, Liu JS. 2004a. Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. Am J Hum Genet 74:495-510.

Kang SJ, Gordon D, Finch SJ. 2004b. What SNP genotyping errors are most costly for genetic association studies? Genet Epidemiol 26:132-141.

Kingman JFC. 1982. The Coalescent. Stochastic Processes and their Applications 13:235-248.

Kirk KM, Cardon LR. 2002. The impact of genotyping error on haplotype reconstruction and frequency estimation. Eur J Hum Genet 10:616-622.

Kraft P, Cox DG, Paynter RA, 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.

Kreienbrock L, Schach S. 2005. Epidemiologische Methoden.

Kuchenhoff H, Mwalili SM, Lesaffre E. 2006. A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX. Biometrics 62:85-96.

Lai R, Zhang H, Yang Y. 2007. Repeated measurement sampling in genetic association analysis with genotyping errors. Genet Epidemiol 31:143-153.

Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. 2003. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 55:56-65.

Lamina C, Steffens M, Mueller J, Lohmussaar E, Meitinger T, Wichmann HE. 2005. Genetic diversity in German and European populations: looking for substructures and genetic patterns. Gesundheitswesen 67 Suppl 1:S127-S131.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la BM, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E,

Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Leal SM. 2005. Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. Genet Epidemiol 29:204-214.

Levenstien MA, 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

Lewontin RC. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49:49-67.

Lincoln SE, Lander ES. 1992. Systematic detection of errors in genetic linkage data. Genomics 14:604-610.

Little R.J.A, Rubin D.B. 2002. Statistical Analysis with Missing Data.

Liu W, Zhao W, Chase GA. 2006. The impact of missing and erroneous genotypes on tagging SNP selection and power of subsequent association tests. Hum Hered 61:31-44.

Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM. 2000. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am J Hum Genet 67:383-394.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356-369.

McCullagh P, Nelder JA. 2008. Generalized Linear Models.

McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH, Cohen JC. 2007. A common allele on chromosome 9 associated with coronary heart disease. Science 316:1488-1491.

Mensah FK, Gilthorpe MS, Davies CF, Keen LJ, Adamson PJ, Roman E, Morgan GJ, Bidwell JL, Law GR. 2007. Haplotype uncertainty in association studies. Genet Epidemiol 31:348-357.

Mitchell AA, Cutler DJ, Chakravarti A. 2003. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. Am J Hum Genet 72:598-610.

Morris RW, Kaplan NL. 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221-233.

Morris RW, Kaplan NL. 2004. Testing for association with a case-parents design in the presence of genotyping errors. Genet Epidemiol 26:142-154.

Morrissey MJ, Spiegelman D. 1999. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. Biometrics 55:338-344.

Morton N. 1982. Outline of genetic epidemiology.

Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC. 2006. Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered 61:55-64.

Moskvina V, Schmidt KM. 2006. Susceptibility of biallelic haplotype and genotype frequencies to genotyping error. Biometrics 62:1116-1123.

Mwalili SM, Lesaffre E, Declerck D. 2005. A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. Journal of the Royal Statistical Society: Series C Applied statistics 54:77-93.

Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157-169.

Nyholt DR. 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765-769.

Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6:847-859.

Quade SR, Elston RC, Goddard KA. 2005. Estimating haplotype frequencies in pooled DNA samples when there is genotyping error. BMC Genet 6:25

Redden DT, Allison DB. 2003. Nonreplication in genetic association studies of obesity and diabetes research. J Nutr 133:3323-3326.

Rice KM, Holmans P. 2003. Allowing for genotyping error in analysis of unmatched case-control studies. Ann Hum Genet 67:165-174.

Ritchie MD, Hahn LW, Moore JH. 2003. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150-157.

Ruckerl R, Greven S, Ljungman P, Aalto P, Antoniades C, Bellander T, Berglind N, Chrysohoou C, Forastiere F, Jacquemin B, von Klot S, Koenig W, Kuchenhoff H, Lanki T, Pekkanen J, Perucci CA, Schneider A, Sunyer J, Peters A. 2007. Air pollution and inflammation (interleukin-6, C-reactive protein, fibrinogen) in myocardial infarction survivors. Environ Health Perspect 115:1072-1080.

Schaid DJ. 2004. Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348-364.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425-434.

Schoenborn V, Heid IM, Vollmert C, Lingenhel A, Adams TD, Hopkins PN, Illig T, Zimmermann R, Zechner R, Hunt SC, Kronenberg F. 2006. The ATGL gene is associated with free fatty acids, triglycerides, and type 2 diabetes. Diabetes 55:1270-1275.

Schunkert H, Gotz A, Braund P, McGinnis R, Tregouet DA, Mangino M, Linsel-Nitschke P, Cambien F, Hengstenberg C, Stark K, Blankenberg S, Tiret L, Ducimetiere P, Keniry A, Ghori MJ, Schreiber S, El Mokhtari NE, Hall AS, Dixon RJ, Goodall AH, Liptau H, Pollard H, Schwarz DF, Hothorn LA, Wichmann HE, Konig IR, Fischer M, Meisinger C, Ouwehand W, Deloukas P, Thompson JR, Erdmann J, Ziegler A, Samani NJ. 2008. Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. Circulation 117:1675-1684.

Seaman SR, Holmans P. 2005. Effect of genotyping error on type-I error rate of affected sib pair studies with genotyped parents. Hum Hered 59:157-164.

Sobel E, Papp JC, Lange K. 2002. Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet 70:496-508.

Spinka C, Carroll RJ, 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.

Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK, Strobbe LJ, Swinkels DW, van Engelenburg KC, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Saez B, Lambea J, Godino J, Polo E, Tres A, Picelli S, Rantala J, Margolin S, Jonsson T, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, Johannsson J, Sveinsson T, Myrdal G, Grimsson HN, Sveinsdottir SG, Alexiusdottir K, Saemundsdottir J, Sigurdsson A, Kostic J, Gudmundsson L, Kristjansson K, Masson G, Fackenthal JD, Adebamowo C, Ogundiran T, Olopade OI, Haiman CA, Lindblom A, Mayordomo JI, Kiemeney LA, Gulcher JR, Rafnar T, Thorsteinsdottir U, Johannsson OT, Kong A, Stefansson K. 2008. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 40:703-706.

Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, Entz P, Suk EK, Toliat MR, Klopp N, Caliebe A, Konig IR, Kohler K, Ludemann J, Diaz LA, Fimmers R, Lichtner P, Ziegler A, Wolf A, Krawczak M, Nurnberg P, Hampe J, Schreiber S, Meitinger T, Wichmann HE, Roeder K, Wienker TF, Baur MP. 2006. SNP-based analysis of genetic substructure in the German population. Hum Hered 62:20-29.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162-1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-989.

Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC. 2003a. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. Human Heredity 55:27-36.

Stram DO, Leigh PC, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. 2003b. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179-190.

Thomas D, Stram D, Dwyer J. 1993. Exposure measurement error: Influence on exposure-disease relationships and methods of correction. Annu Rev Public Health 14:69-93.

Thomas DC. 2004. Statistical Methods in Genetic Epidemiology.

Tintle NL, Gordon D, McMahon FJ, Finch SJ. 2007. Using duplicate genotyped data in genetic analyses: testing association and estimating error rates. Stat Appl Genet Mol Biol 6:Article4

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di F, V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu

X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M. 2001. The sequence of the human genome. Science 291:1304-1351.

Wacholder S, Armstrong B, Hartge P. 1993. Validation studies using an alloyed gold standard. Am J Epidemiol 137:1251-1258.

Wichmann HE, Gieger C, Illig T. 2005. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67 Suppl 1:S26-S30.

Wong MY, Day NE, Luan JA, Wareham NJ. 2004. Estimation of magnitude in gene-environment interactions in the presence of measurement error. Stat Med 23:987-998.

Xu H, Wu X, Spitz MR, Shete S. 2004. Comparison of haplotype inference methods using genotypic data from unrelated individuals. Hum Hered 58:63-68.

Ye S, Willeit J, Kronenberg F, Xu Q, Kiechl S. 2008. Association of genetic variation on chromosome 9p21 with susceptibility and progression of atherosclerosis: a population-based, prospective study. J Am Coll Cardiol 52:378-384.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79-91.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316:1336-1341.

Zhao LP, Li SS, 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 WS, Fung WK, Guo J. 2007. Incorporating genotyping uncertainty in haplotype frequency estimation in pedigree studies. Hum Hered 64:172-181.

Zou G, Pan D, Zhao H. 2003. Genotyping error detection through tightly linked markers. Genetics 164:1161-1173.

Zou G, Zhao H. 2003. Haplotype frequency estimation in the presence of genotyping errors. Hum Hered 56:131-138.

A8 Curriculum Vitae

Name:	Claudia Lamina
Date, Place of Birth:	March 13th, 1977 in Augsburg, Germany

Education:

06/1996:	Abitur at the Rudolf-Diesel-Gymnasium, Augsburg
10/1996 – 05/2003	University Diploma in Statistics (Area of Application: Biometrics), University of Munich (LMU), München, Germany
08/2000 – 05/2001:	Exchange student in Statistics at the Kansas State University, Manhattan, KS, USA (Contact Scholarship of the LMU with partner universities)

Professional Experience:

11/1998 – 07/2000:	Student Assistant at the Institute for Medical Statistics and Epidemiology (IMSE), Technical Unversity Munich
01/2001 – 05/2001:	Student Statistical Consultant at the Kansas State University, Manhattan, KS, USA
07/2001 – 12/2002:	Student Assistant at the Sylvia Lawry Centre for Multiple Sclerosis Research, Munich
07/2003 – present:	Research Assistant at the Institute of Epidemiology at the Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, and PhD Student at the Ludwig-Maximilian-Universität München