Case-Control Association Tests Correcting for Population Stratification

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch–Naturwissenschaftlichen Fakultäten
der Georg–August–Universität zu Göttingen

vorgelegt von
Karola Köhler
aus
Göttingen

Göttingen 2005

D7
Referent: Prof. Dr. Manfred Denker
Korreferentin: Prof. Dr. Heike Bickeböller
Tag der mündlichen Prüfung: 25.01.2006

Danksagung

Ganz besonders möchte ich mich bei Frau Prof. Dr. Heike Bickeböller bedanken, die mir das interessante Thema vorgeschlagen und mich während der Entstehung der Arbeit maßgeblich begleitet hat. Sie hat es mir ermöglicht, an vielen Workshops und Tagungen teilzunehmen, meine Arbeit vorzustellen und neue Anregungen zu bekommen. Außerdem danke ich Herrn Prof. Dr. Manfred Denker für die Übernahme des Erstgutachtens.
Weiterhin möchte ich mich bei Herrn Prof. Dr. Jonathan K. Pritchard dafür bedanken, mir den Quellcode seiner Programme STRUCTURE und STRAT sowie sein eigenes Simulationsprogramm zur Verfügung gestellt zu haben. Ein besonderes Dankeschön gilt auch Dr. med. Michael Steffens für die Diskussionen über die Genomic Control Studie sowie Melanie Bergmann für das sehr sorgfältige und zeitaufwendige Korrekturlesen meiner Arbeit. Außerdem bedanke ich mich bei meinen Kolleginnen und Kollegen der Abteilungen Genetische Epidemiologie und Medizinische Statisitik, die mir immer geholfen haben, offene Fragen zu klären.
Nicht vergessen zu erwähnen möchte ich meine Familie und meine Freunde, die jederzeit eine große moralische Unterstützung für mich waren.

Die Arbeit wurde teilweise durch das Bundesministerium für Bildung und Forschung (BMBF) im Rahmen der genetisch-epidemiologischen Methodenzentren im nationalen Genomforschungsnetz gefördert (Fördernummern: 01GS0204, 01GR0462, 01GS0422).

Contents

1 Introduction

In a case-control study a group of subjects affected with a disease (cases) is compared to a group of unaffected subjects (controls) with respect to potential exposures of risk factors. In genetic case-control association studies the exposures of primary interest are genetic factors. In the candidate gene approach so called candidate genes expected to be functionally related to disease causing pathways are chosen to be investigated. Each study participant is genotyped at one or a few genetic marker loci which are specific positions on the chromosomes lying directly on or very close to the candidate genes. As the genotyping result the genotype consisting of two alleles at the marker locus is determined for each individual. Most often these marker loci are single nucleotide polymorphisms which have two possible alleles and three possible genotypes in the population. The genotype distribution at a marker locus is investigated with respect to differences between cases and controls. In case of such a difference marker locus and disease are said to be associated. The simplest test for association between the alleles at the marker locus and the disease status is Pearson's χ^2-test for a 2×2 table. However, there are different reasons for an association. The first reason is that the marker locus is itself a locus directly related to the disease, thus the association is causal. An association also can be observed if marker and disease locus are in close proximity on the genome due to linkage disequilibrium caused by the tendency of certain alleles appearing together on short chromosomal segments.
However, an association can also be observed due to unobserved population stratification. Li (1969) first noted the possible importance of unobserved population structure for genetic association studies. Population structure may lead to spurious associations if the allele distribution is different between the subpopulations and if the general disease risk varies between the subpopulations. Under these conditions population stratification is said to act as a confounder in the case-control association study if the general epidemiological terminology is used. If population stratification is not accounted for the number of false positive association tests increases. In the middle of the nineties with improvements in genotyping technology the concerns about population stratification became more relevant (e.g. Lander and Schork, 1994). Furthermore, it turned out that association analysis would become an important tool to find the genetic basis of complex diseases (e.g. Risch and Merikangas, 1996). Most of the common diseases are complex diseases where there clearly is a genetic basis but the genetic model is not clear. Many genes with small genetic effects are expected to contribute to a complex disease but the knowledge about complex diseases is still rather small. To overcome the problems of population stratification Spielman et al. (1993) introduced the TDT (transmission disequilibrium test) as one of the first family based association tests and Ewens and Spielman (1995) explicitly showed that this test for association is robust against population structure. Case-parents trios consisting of one affected subject and its parents have to be recruited and the transmission of alleles

from the parents to the affected offspring is investigated. Subsequently, the number of family-based association tests has been grown rapidly including sibs of the proband and allowing for missing parental information (e.g. Spielman and Ewens, 1998; Knapp, 1999). Rabinowitz and Laird (2000) developed FBAT as a unified approach for the analysis of family based association studies including pedigrees with arbitrary structure and arbitrary missing marker information. However, there are also a lot of convincing arguments against the use of family-based association tests (e.g Risch and Teng, 1998). The most important argument is that case-control studies are more easy to be implemented than family-based studies. For complex diseases large samples are required to identify genes with small effects. However, it is difficult and expensive to collect a large number of families. For late-onset diseases it is even impossible to collect parents of the affected subjects.

As an alternative to family-based association tests Devlin and Roeder (1999) proposed to continue with case-control studies but to correct them for population stratification. The idea is to genotype additional genetic markers which are not associated with the disease to make inference about the population structure in the sample. Subsequently, several methods were proposed to account for population structure in case-control studies. These methods broadly follow one of two concepts: Genomic Control (GC) or Structured Association (SA). Genomic Control tests empirically estimate the variance inflation of the original test statistic (Devlin and Roeder, 1999; Reich and Goldstein, 2001). In the model based Structured Association approach (Pritchard et al., 2000a,b; Satten et al., 2001; Zhu et al., 2002; Chen et al., 2003; Hoggart et al., 2003; Purcell and Sham, 2004) population structure is directly inferred and the test of association incorporates the estimated population structure.

The main topic of the thesis is to introduce a new method of Structured Association. Furthermore different case-control association tests correcting for population stratification are investigated and the new method is compared to existing methods of Structured Association as well as to Genomic Control.

The thesis starts with two introductory chapters, chapter 2 about population genetics and chapter 3 about genetic case-control association studies to introduce the basic principles which are necessary to understand the concepts of Genomic Control and Structured Association.

In chapter 4 the method of Genomic Control is introduced. We theoretically investigate Genomic Control and present some new results about the performance of Genomic Control. Here we concentrate on the variability in the estimation of the variance inflation of the test statistic and its impact on the type-I error rate of Genomic Control.

The main focus of the thesis is chapter 5 on Structured Association. There are several Structured Association approaches proposed in the literature which can be mainly divided into two categories as one- or two-step approaches. Pritchard et al. (2000a,b) proposed the two-step approach where the first step consists of modelling population structure for

the entire study sample and the second step is the test of association based on the inferred structure. In contrast, the idea of Satten et al. (2001) was to simultaneously estimate population structure and test for association. Although most of the Structured Association methods have in common that they use a probability based approach to split the entire sample into subpopulations they differ in other aspects: Pritchard et al. (2000a) use a Bayesian model for population structure where it is possible to include admixed individuals with alleles from more than one subpopulation. The approach of Satten et al. (2001) is based on a mixture model to split the entire sample into discrete subpopulations. The association test of Pritchard et al. (2000b) is a likelihood ratio test whereas Satten et al. (2001) apply a logistic regression model to test for association. More recently further Structured Association approaches have been developed. Hoggart et al. (2003) proposed a one-step approach where population structure is modelled similar to Pritchard et al. (2000a) in a Bayesian framework but for the test of association a logistic regression model is fitted similar to Satten et al. (2001). Purcell and Sham (2004) introduced a simpler two-step approach like Pritchard et al. (2000a,b) but used the EM algorithm to infer population structure in a mixture model like Satten et al. (2001).

Thus, there are a lot of Structured Association approaches proposed in the literature but no systematic attempt has been made to theoretically investigate and empirically compare some of these. Furthermore most of these approaches are rather complicated and it is not clear how well these behave in practice. Thus, the main subject of this thesis is to propose a new and rather simple method of Structured Association and systematically investigate some variations of this method in order to analyze the respective influence of different clustering approaches as well as that of different test statistics. Our approach combines some aspects of Pritchard et al. (2000a,b), Satten et al. (2001) and Purcell and Sham (2004). Like the method of Purcell and Sham (2004) it is a two-step approach based on a mixture model for discrete subpopulations. The most important difference to Pritchard et al. (2000a) and Purcell and Sham (2004) is that our clustering method incorporates the information about the disease status for identifying subpopulations. In this respect, our approach is similar to the simultaneous approach of Satten et al. (2001) who propose to infer population structure not only conditionally on the phenotypic information but also conditionally on the candidate gene. We show here that even in a two-step approach it is necessary to include the information about the phenotype if the association test is based on the likelihood for the genotype data at the candidate locus given the phenotype data. Otherwise the estimated subpopulation proportions in the case and the control group are biased leading to an inflated type-I error rate. The most popular Structured Association approach as proposed by Pritchard et al. (2000a,b) violates this principle and is thus expected to lead to an inflated type-I error rate.

For the association test we propose a new Wald test statistic which could be applied in a two-step approach instead of the likelihood ratio test of Pritchard et al. (2000b). The

Wald test averages the estimated allele frequency differences between cases and controls over all subpopulations. In contrast to the likelihood ratio test, the Wald test only has 1 degree of freedom and is theoretically designed for situations where no interaction between the population structure and the effects of the candidate gene is expected.
In addition to describing the new Structured Association method we give a systematic overview over other methods of Structured Association also explaining a few new ideas how to extend some of these. Moreover, a theoretical comparison of our method to the other methods of Structured Association is included.
In chapter 6 the impact of population stratification on case-control association studies is investigated in realistic situations of only small to moderate population stratification as within Europe or even within Germany. There has been a lot of debate within the scientific community about the impact of population stratification (e.g. Wacholder et al., 2002; Thomas and Witte, 2002). Very recently, the opinion seems to prevail that even in populations that seem to be rather homogeneous there can be a measurable impact of hidden population stratification on the association results (Marchini et al., 2004; Freedman et al., 2004; Campbell et al., 2005). To measure the degree of population stratification within Germany a Genomic Control study was conducted within the framework of the German National Genome Research Network (NGFN = Nationales Genomforschungsnetz). The study was mainly analyzed by the Genetic Epidemiological Center in Bonn but we also contributed to the analysis by proposing and calculating the prediction rate as a new measure for the prediction of population membership. Here the study is briefly described and the main results are summarized with an emphasis on the prediction rate. It turns out that there is a measurable difference between North and South Germany, but the difference is too small to be identified by a probability based clustering algorithm. Thus, within Germany methods of Structured Association cannot successfully be applied for correcting case-control association tests.
Our theoretical results about Structured Association and Genomic Control are verified in a large simulation study comparing our new method of Structured Association to Structured Association methods similar to the method of Pritchard et al. (2000a,b) and to Genomic Control (see chapter 7). Data sets are simulated for realistic situations of large case-control studies with only small to moderate amount of population stratification as expected between European populations. We compare our results to previously published simulations (Bacanu et al., 2000; Pritchard and Donnelly, 2001; Devlin et al., 2001a; Chen et al., 2003; Marchini et al., 2004; Shmulewitz et al., 2004) regarding Genomic Control or the Structured Association method proposed by Pritchard et al. (2000a,b). We can conclude from our results that the Structured Association method we propose is most often superior to the other Structured Association methods investigated in the simulations. A disadvantage of Genomic Control turns out to be the large variation in estimating the variance inflation factor as well as the power loss if population structure increases. Alto-

gether, the simulations show that the model based approach of Structured Association if applied correctly is in general superior to the Genomic Control approach. This holds at least in situations of rather simple population structure as investigated in the simulations. The results of the main chapters are discussed at the end of each of these chapters. In section 5.4 a theoretical and more technical discussion about the different Structured Association methods can be found. Section 6.3 contains a small discussion about the impact of population stratification on case-control studies in realistic situations of small to moderate population stratification. The simulation results and the different performance of Structured Association and Genomic Control are discussed in detail in section 7.3. Thus, the last chapter 8 only contains a short summary and an outlook what has to be investigated further.
Finally, the notation and the statistical theory applied for the derivation of the methods are summarized in the appendix A.

2 Population genetics

This chapter gives an overview about the basic principles of population genetics necessary to understand the concepts of correcting case-control association studies with respect to population stratification. The chapter starts with an introduction to genetics. Subsequently, basic concepts of population genetics are described, first for random mating populations and later on for subdivided populations. The chapter finishes with deriving some methods for statistical inference in subdivided populations.

2.1 Introduction to Genetics

The basic concept of genetics that human characteristics are inherited from parents to offspring in discrete units called genes is well known. However, to use statistics in human genetics, the statistician has to be familiar with some principles of genetics as well as some genetic terminology which is described here mainly based on Sham (1998).
First of all the physical and chemical structure of a gene shall be described. The *chromosomes* are the physical location of the genes in the cell nucleus, the *DNA* (deoxyribonucleic acid) is the chemical structure of the chromosomes carrying the genetic information. Each chromosome contains two very long strands of DNA which are normally bound to each other and twisted around each other as a double helix. One strand of the DNA consists of a sequence of *nucleotides* which mainly differ in the *nitrogenous base* belonging to the nucleotide. There are four different bases called adenine (A), guanine (G), cytosine (C) and thymine (T). The other strand of the DNA is complementary in sequence where A is always paired with T and G with C. The genetic information is contained in the sequence of the nucleotides of one strand. This information has to be translated into *protein molecules* which perform all kinds of structural and biochemical functions. Each protein molecule is a chain of *amino acids*. These amino acids exist in twenty different forms. Each possible triplet of nucleotides of the DNA represents a special amino acid. Most of the amino acids are coded by different triplets since there are 64 possible triplets. Based on this knowledge, a more precise definition of a *gene* is possible: a gene is a segment of DNA within a chromosome that specifies the amino acid sequence of a single subunit of a protein.
The knowledge about the inheritance of chromosomes is also crucial. The total genetic information of an individual is contained in 23 pairs of chromosomes including 22 pairs of *autosomes* and 2 sex chromosomes. The two chromosomes of a chromosome pair are called *homologous*. A set of these 23 chromosome pairs is contained in the nucleus of each cell and hence duplicated during normal cell division called *mitosis*. All the cells of an individual are ultimately derived from a single cell called the *zygote* which is formed by the union of two *gametes*, one from each parent. Each gamete contributes a *haploid* (single) set of 23 chromosomes so that the zygote receives a *diploid* (double) set of 23 pairs

of chromosomes. Gametes are produced by a special form of cell division called *meiosis*. Meiosis involves the reshuffling of genetic material by the exchange of chromosome segments between the two homologous chromosomes. Hence the chromosomes in the gamete consist of alternating segments of paternally and maternally inherited DNA. Thus, each chromosome pair of the offspring consists of a maternally and paternally inherited chromosome which is composed of a set of "chunks" which are inherited as unbroken unit from the parent. The "chunks" consist of alternating sequences of maternally and paternally inherited DNA of the parent. Thus, in the last sentence maternally and paternally refers to grandmother and grandfather if the perspective of the offspring is chosen. The tendency of short chromosomal segments to be inherited intact from parent to offspring is known as genetic *linkage*.

The human genome project, 1990-2003, (Human Genome Management Information System, 2003) provided some important knowledge about the *human genome* which is the scientific term for the complete set of the human DNA. In 2001, the first analysis of the working draft human genome sequence was published (McPherson et al., 2001) and in 2003 a reference sequence of the human genome was completed marking the end of the human genome project. Sequencing showed that the human genome contains approximately 3 billion nucleotide bases. However, genes only comprise about 2% of the human genome and the remaining part is non-coding. The number of genes initially was overestimated, recent estimates assume that the human genome consists of 20,000 -25,000 genes (Stein, 2004).

The human genome sequence is almost (99.9%) identical in all people. Nevertheless changes in the DNA occur from time to time and such *mutations* introduce diversity in the population. Variations in non-coding DNA usually have no observable effect. Mutations in coding regions sometimes also have no effect if the new triplet codes for the same amino acid as the original. However, such mutations often cause a change in the amino acid sequence. Sometimes the resulting protein has similar properties as the original but one mutation can also be responsible for a major disorder if a harmful protein is produced. The presence of different DNA sequences at the same position in a population is known as a genetic *polymorphism* if all sequences are occurring more frequently than can be accounted for by mutation alone. One of the most common types of sequence variation is a *single nucleotide polymorphism* (*SNP*) where individuals differ in their DNA sequence only in one single base. The number of single nucleotide polymorphisms in the human genome is estimated at least at ten million. About 3 million of these have already been identified and are recorded in SNP databases, as for example dbSNP of the NCBI (National Center for Biotechnology Information).

To derive statistical models for genetics some genetic terminology has to be introduced. A *locus* is defined as a specific position in the genome. *Alleles* are the alternative DNA sequences at a locus. The two alleles at the same locus of one individual are defined as his

or her *genotype*. If the two alleles are identical the genotype is said to be *homozygous*, otherwise it is called *heterozygous*. For example, single nucleotide polymorphisms usually are *diallelic* loci with only two possible alleles and three possible genotypes.
If only a single locus is considered and rare exceptions are disregarded the *law of segregation* holds: during reproduction the parents give with equal probability one of the two alleles from a specific locus to their offspring, independently from the other parent. If there is no parental information available it is unknown for the offspring which allele is inherited from the mother and which one from the father.

2.2 Random mating populations

In this section basic principles of population genetics for random mating are introduced as described in Gillespie (1998) and Sham (1998). The most idealistic population model is a random mating population of infinite size. Random mating describes the situation where mating is done between randomly chosen individuals.
The notation introduced in this chapter is applied consistently throughout this thesis. We consider a single diallelic autosomal marker locus l with two alleles B and b. For an individual $i \in \mathbb{N}$ in the population such a locus can be described by two Bernoulli distributed random variables X_{ilj} for the two alleles at the two DNA-strands $j = 1, 2$ at the same locus. The random variables X_{ilj} take the values 1 if allele B is present and 0 if allele b is present. The probability of choosing randomly allele B from the population is defined as $\varphi_l = P(B) = P(X_{ilj} = 1)$. In genetics, it is common to denote φ_l as allele frequency, although statistically speaking it is a probability. Since this term is common we also want to use it here. The genotype of individual i is then uniquely identified by the sum of the two alleles $X_{il} := X_{il1} + X_{il2}$. The number 0 denotes the homozygous genotype bb, the number 1 the heterozygous genotype Bb and the number 2 the homozygous genotype BB. For the heterozygous genotype Bb the order generally is not meaningful since it is usually unknown which allele is inherited from the mother and which one from the father.

2.2.1 Hardy-Weinberg equilibrium

The first milestone in population genetics was the discovery of the simple Hardy-Weinberg law which is valid in the equilibrium state of a random mating population of infinite size. It describes the relationship between allele and genotype probabilities at a fixed autosomal locus l. The Hardy-Weinberg law says that in the equilibrium state the genotype frequencies can be obtained from allele frequencies by

$$P(X_{il} = s) = \binom{2}{s} \varphi_l^s (1 - \varphi_l)^{2-s}$$

for $s = 0, 1, 2$. In other words in *Hardy-Weinberg equilibrium* (*HWE*) the genotypes are $B(2, \varphi_l)$ -distributed. This law simply uses that both alleles X_{il1} and X_{il2} are randomly

chosen under the assumption of random mating and independently drawn from the two parents. Hence the equilibrium state is already reached after one generation of random mating in an infinite population. Only if allele frequencies in the original population are different for both sexes it takes one generation of random mating to reach equal allele frequencies for both sexes and Hardy-Weinberg equilibrium is only reached after two generations of random mating. However, in finite populations *genetic drift* can cause random changes in allele frequencies from one generation to the other eventually leading to a decay of heterozygosity and removing genetic variation. Since genetic drift is a very weak evolutionary force in large populations significant deviations from Hardy-Weinberg equilibrium due to genetic drift are only expected in very small populations. Moreover, there are also other evolutionary forces such as mutation, migration and selection which could destroy Hardy-Weinberg equilibrium (e.g. Maynard Smith, 1989).

2.2.2 Linkage equilibrium

Linkage equilibrium is an often desired property between two different loci l and l' with alleles B, b as well as B', b'. The *multilocus genotype* of each individual consists of two genotypes $X_{il} = X_{il1} + X_{il2}$ and $X_{il'} = X_{il'1} + X_{il'2}$ each composed of two Bernoulli distributed alleles with allele frequency $P(B)$ for locus l and $P(B')$ for locus l'. The alleles on the same DNA strand are called a *haplotype*, i.e. the two haplotypes are $(X_{il1}, X_{il'1})$ and $(X_{il2}, X_{il'2})$. The haplotype concept can be extended to more than two loci and a haplotype in general denotes all the alleles from the same gamete. The multilocus genotype of each individual consists of two haplotypes, one inherited from the father and the other inherited from the mother. However, as mentioned before, usually only multilocus genotype data are available because laboratory methods routinely only measure genotypes. Haplotypes are normally unknown since it is very expensive to determine which alleles are on the same DNA strand. The frequencies of the four possible haplotypes are denoted as $P(BB')$, $P(Bb')$, $P(bB')$ and $P(bb')$. The two loci are said to be in *linkage equilibrium* if the two alleles X_{ilj} and $X_{il'j}$ of the same haplotype are independently Bernoulli distributed, thus the haplotype frequencies are the product of the corresponding allele frequencies, e.g. $P(BB') = P(B)P(B')$. The deviation of the frequency $P(BB')$ from its equilibrium value is called *linkage disequilibrium* (LD) $D_{ll'} = P(BB') - P(B)P(B')$. The linkage disequilibrium can also be defined as the covariance between the two alleles of the same strand $D_{ll'} = \mathrm{Cov}\,(X_{ilj}, X_{il'j})$ because

$$D_{ll'} = \mathrm{Cov}\,(X_{ilj}, X_{il'j}) = P(X_{ilj} = 1, X_{il'j} = 1) - P(X_{ilj} = 1)P(X_{il'j} = 1).$$

There are different ways to standardize measures for linkage disequilibrium. One idea is to take the correlation coefficient

$$\Delta_{ll'} = \mathrm{Corr}\,(X_{ilj}, X_{il'j})$$

to describe linkage disequilibrium. If two loci are in linkage disequilibrium there is said to be an *allelic association* between these two loci.

2.2.3 Linkage disequilibrium due to tight linkage

Linkage disequilibrium often exists between loci lying close to each other on the genome. In this case the LD is caused by genetic linkage (see section 2.1). As a result, haplotypes on these segments may be preserved over a large number of generations. The recombination fraction $\theta_{ll'}$ between the two loci measures the extent of linkage. A gamete produced by an individual is said to be *non-recombinant* with respect to two loci if it contains the haplotype of one of the parental gametes. In contrast, *recombinant* gametes contain a new combination of the two alleles at the two loci, one from the paternal and one from the maternal gamete. The *recombination fraction* $\theta_{ll'}$ between the two loci is defined as the probability that a gamete is recombinant. Thus, two loci on different chromosomes have a recombination fraction of $\theta_{ll'} = 0.5$. The recombination fraction becomes the smaller the more tightly linked the loci are because the probability that the chromosomal segment covering both loci is inherited intact from parent to offspring increases with close proximity on the genome. The maintenance of LD over generations by tight linkage in an infinite random mating population is dependent on the recombination fraction in the following way

$$D_{ll'}^{(t)} = (1 - \theta_{ll'})^t D_{ll'}^{(0)}$$

where $D_{ll'}^{(t)}$ denotes the LD after t generations and hence $D_{ll'}^{(0)}$ the initial LD. The derivation of the formula is shown in Sham (1998), for example. The formula shows that for $\theta_{ll'} > 0$ the ultimate state for $t \longrightarrow \infty$ is linkage equilibrium, but the decay of linkage disequilibrium between closely linked markers is very slow. If there is initial LD, even between unlinked loci ($\theta_{ll'} = 0.5$) it lasts some generations until a state close to linkage equilibrium is reached. Furthermore, in natural populations the decay of LD is opposed by several evolutionary forces which could increase LD, e.g. random genetic drift and mutations in finite populations. Thus, in most human populations linkage disequilibrium can be observed between tightly linked markers with a recombination fraction close to zero, up to 50 kb (50 kilo bases; 50,000 bases), occasionally also up to 500 kb (Abecasis et al., 2001). For loci located not so close to each other linkage equilibrium can be assumed.

2.2.4 Extension to multiple alleles

The concepts of population genetics can be extended to loci which have more than two alleles. The most often used genetic markers with multiple alleles are *microsatellites*. Microsatellite loci often cover some hundreds of nucleotides. The variation between individuals occurs in the form of a variable number of repeats of a particular sequence of base

pairs. The repeated sequence usually is very short (2 to 4 base pairs). Alleles are uniquely identified by counting the number of repeats. However, genotyping of microsatellites is much more expensive and error-prone than genotyping of single nucleotide polymorphisms. It is not expected that microsatellites play a major role in the future, thus we concentrate our work on diallelic marker. However, we briefly would like to mention how to extent the concepts to multiallelic markers.
Let l be a multiallelic marker locus with the alleles $B^{(1)}, \ldots, B^{(R_l)}$. For an individual i such a locus can be described by two random vectors $\boldsymbol{X}_{ilj} = (X_{ilj}^{(1)}, \ldots, X_{ilj}^{(R_l)})'$ for both strands $j = 1, 2$. The entry $X_{ilj}^{(r)}$ takes the value 1, if the allele on strand j is allele $B^{(r)}$ and value 0 otherwise. Thus, $X_{ilj}^{(r)}$ is Bernoulli($\varphi_l^{(r)}$)-distributed where $\varphi_l^{(r)}$ is the frequency of allele $B^{(r)}$. Since the components of $\boldsymbol{X}_{ilj}$ sum up to 1, the vector $\boldsymbol{X}_{ilj}$ is multinomial$(1, \boldsymbol{\varphi}_l)$-distributed with $\boldsymbol{\varphi}_l = (\varphi_l^{(1)}, \ldots, \varphi_l^{(R_l)})'$. The basic definitions are extended considering each component of the vector $\boldsymbol{X}_{ilj}$ separately. A locus is in Hardy-Weinberg equilibrium if each component of the vector $\boldsymbol{X}_{il1}$ for strand 1 and each component of the vector $\boldsymbol{X}_{il2}$ for strand 2 are independent. Thus, if Hardy-Weinberg equilibrium exists the vector of genotypes $\boldsymbol{X}_{il} = (X_{il}^{(1)}, \ldots, X_{il}^{(R_l)})'$ where $X_{il}^{(r)} = X_{il1}^{(r)} + X_{il2}^{(r)}$ is multinomial$(2, \boldsymbol{\varphi}_l)$-distributed. Furthermore, two loci l and l' are in linkage equilibrium if each component of the vector $\boldsymbol{X}_{ilj}$ for locus l is independent from each component of $\boldsymbol{X}_{il'j}$ for locus l' and the same strand j. To model linkage disequilibrium between the two loci $R_l R_{l'}$ covariances $D_{ll'}^{(rr')} = \mathrm{Cov}\,(X_{ilj}^{(r)}, X_{il'j}^{(r')})$ have to be considered which can be summarized in different ways to a single measure of linkage disequilibrium.

2.3 Models for structured populations

Most natural populations deviate in some way from random mating because they are not homogeneous but structured in some form. The main focus of this section is on the simplest form of population structure where a population consists of several discrete subpopulations. The classical concepts of populations genetics for subdivided populations can be found in Excoffier (2000). Additionally, we also introduce some recently developed concepts and show how these are related to the classical definitions. However, before such a subdivided population is considered the general concept of inbreeding is introduced.

2.3.1 The general concept of inbreeding and allelic correlations

Inbreeding is one important reason for a departure from random mating. Inbreeding occurs when individuals are more likely to mate with relatives than with randomly chosen subjects. In this context the term relative is used in a broad sense, i.e. relatives can also be individuals from the same village or the same region distantly related to each other. In this case the two alleles X_{il1} and X_{il2} at the same locus l of individual i are

not independent anymore. The *inbreeding coefficient* (Wright, 1922) denoted by F_i is defined as the correlation between these two alleles, i.e. $F_i = \mathrm{Corr}\,(X_{il1}, X_{il2})$. Thus, F_i measures the correlation between uniting gametes. Under the assumption that the inbreeding coefficient is the same for all individuals in the population, i.e. $F_i = F$ a formula for the frequency of genotype BB in the population can be directly derived by using the definition of the correlation

$$F = \mathrm{Corr}\,(X_{il1}, X_{il2}) = \frac{\mathrm{Cov}\,(X_{il1}, X_{il2})}{\sqrt{\mathrm{Var}\,(X_{il1})\mathrm{Var}\,(X_{il2})}} = \frac{\mathrm{E}\,(X_{il1}X_{il2}) - \varphi_l^2}{\varphi_l(1-\varphi_l)} = \frac{P(BB) - \varphi_l^2}{\varphi_l(1-\varphi_l)}.$$

The probabilities for all three genotypes follow immediately as given here

$$\begin{aligned} P(BB) &= \varphi_l^2(1-F) + \varphi_l F \\ P(Bb) &= 2\varphi_l(1-\varphi_l)(1-F) \\ P(bb) &= (1-\varphi_l)^2(1-F) + (1-\varphi_l)F. \end{aligned}$$

If inbreeding exists in the population the inbreeding coefficient F is always positive ($0 < F \leq 1$) leading to an excess of homozygotes compared to the Hardy-Weinberg equilibrium. F may also be interpreted as the probability of the two alleles being *identical by descent* (*IBD*) which means that the two alleles are descended from the same ancestral allele somewhere in the past. This interpretation follows from the first equation because F is multiplied with the probability φ_l that one of the two alleles is B and the other allele automatically is B due to identity by descent.
It should be pointed out that the definition of the inbreeding coefficient is independent of the concrete locus. Thus, these definitions implicitly assume that the correlations are constant over the whole genome. This is a common assumption because there are no evident biological reasons why the correlations F should vary over the genome.
Allelic correlations are not only restricted to the two alleles of one individual but extend across related individuals. The *kinship coefficient* $f_{ii'}$ is defined as the correlation between an allele selected randomly from individual i and another allele selected randomly from i' from the same locus, i.e. $f_{ii'} = \mathrm{Corr}\,(X_{ilj}, X_{i'lj'})$ for $j, j' \in \{1, 2\}$. The kinship coefficient can also be interpreted as the probability of the alleles from the two individuals being ibd. If the inbreeding coefficients F_i and the kinship coefficients $f_{ii'}$ are given for all individuals i, i' in the population, population structure can be modelled in a very general way allowing for a special relationship between all pairs of individuals. Thus, any form of *cryptic relatedness* is included in the population model. The model is based on a global allele frequency φ_l for each locus with respect to the total population and similarities between individuals are modelled over correlations between alleles from closely related individuals. Thus, the model is referred to as *correlation model* for general population structure.

2.3.2 Theoretical models for discrete subpopulations

A simple model of population structure is that the population is divided into discrete subpopulations $k = 1, \ldots, K$. We would like to consider a model where random mating within the subpopulations is assumed and discuss possible extensions of the model if inbreeding within the subpopulations is allowed. Furthermore we want to investigate the consequences for the total population if two loci are assumed to be in linkage equilibrium within the subpopulations. To model discrete subpopulations based on genetic information different statistical models are introduced. In the literature these models are often not precisely formulated and separated from each other. Instead, usually the model is applied which seems appropriate for deriving a special statistical method. So we make our own attempt of precisely defining models for discrete subpopulations and comparing their statistical properties.

The correlation model for discrete subpopulations The first model is the correlation model, for example applied by Devlin and Roeder (1999). Regarding a fixed locus l only a single allele frequency φ_l for the total population is given. Differences between the subpopulations result from correlations between alleles from members of the same subpopulation. Hence the discrete subpopulations are modelled as a special form of inbreeding. As proposed by Wright (1951) two different kinds of correlations F relative to the total population are distinguished: F_{IT} denotes the *global inbreeding coefficient* with respect to the total population, i.e. the correlation of the two alleles within an individual (I) relative to the total population (T). The *fixation index* F_{ST} describes the correlation between alleles from the same locus of two individuals from the same subpopulation (S) relative to the total population (T). In other words, F_{ST} is in principle the coefficient of kinship including the additional assumption that the correlation is the same between all alleles from individuals of the same subpopulation. Under the assumption of random mating within the subpopulations the equality $F_{IT} = F_{ST}$ holds because two alleles from one individual as well as from the same subpopulation independently and randomly derive from the previous generation within the subpopulation. Thus, in this case it is sufficient to specify the fixation index F_{ST}. Where unambiguous, F_{ST} is abbreviated by F. Thus, population stratification can be modelled by a positive fixation index F_{ST} and like inbreeding population stratification leads to a deviation from Hardy-Weinberg equilibrium for the total population with an excess of homozygotes.

The subpopulation model with fixed subpopulation allele frequencies In this traditional approach fixed allele frequencies φ_{kl} in the subpopulations $k = 1, \ldots, K$ are introduced. Some further notation has to be introduced. The vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ contains the proportion of each subpopulation with respect to the total population, i.e.

the subpopulation proportions sum up to 1 for $k = 1, \ldots, K$. For an individual i in the population the variable Z_i contains the subpopulation where the individual i is from. Thus, it holds that $Z_i = k$ if individual i is from subpopulation k.
As before we assume random mating in the subpopulations. Thus, within the subpopulations the alleles at the same locus from two individuals are independent. This is equivalent to the assumption of Hardy-Weinberg equilibrium in the subpopulations. To model local inbreeding within the subpopulation according to Wright (1951) a *local inbreeding coefficient* F_{IS} relative to the subpopulation (S) could be introduced. We assume that F_{IS} is equal to zero which is equivalent to $F_{IT} = F_{ST}$.
For each locus an overall allele frequency can be defined by averaging over the subpopulation allele frequencies as $\varphi_l := \sum_{k=1}^{K} \pi_k \varphi_{kl}$. Thus, if X_{ilj} is a randomly drawn allele from the total population the expectation can be calculated as

$$\mathrm{E}\,(X_{ilj}) = P(X_{ilj} = 1) = \sum_{k=1}^{K} \pi_k P(X_{ilj} = 1 | Z_i = k) = \varphi_l$$

However, if an individual is randomly drawn from a given subpopulation k the expectation $\mathrm{E}\,(X_{ilj}|Z_i = k) = \varphi_{kl}$ is the subpopulation allele frequency.
In such a model each locus has to be considered separately with its own F_{ST} as can be seen from the following considerations. If we apply the definition of F_{ST} for locus l from the correlation model to this model we can derive the classical relationship

$$F_{ST} = \frac{\mathrm{Var}\,\varphi_{kl}}{\varphi_l(1 - \varphi_l)}\,.$$

According to the definition of F_{ST} as correlation between arbitrarily chosen alleles $(j, j' = 1, 2)$ from two individuals i and i' from the same subpopulation it follows under the assumption of HWE in the subpopulations

$$\begin{aligned}
F_{ST}\varphi_l(1 - \varphi_l) &= \mathrm{Cov}\,(X_{ilj}, X_{i'lj'}) = \mathrm{E}\,(X_{ilj}X_{i'lj'}) - \mathrm{E}\,X_{ilj}\mathrm{E}\,X_{i'lj'} \\
&= \sum_{k=1}^{K} \pi_k P(X_{ilj} = 1, X_{i'lj'} = 1 | Z_i = Z_{i'} = k) - \left(\sum_{k=1}^{K} \pi_k \varphi_{kl}\right)^2 \\
&= \sum_{k=1}^{K} \pi_k \varphi_{kl}^2 - \left(\sum_{k=1}^{K} \pi_k \varphi_{kl}\right)^2 \\
&= \sum_{k=1}^{K} \pi_k (\varphi_{kl} - \varphi_l)^2 =: \mathrm{Var}\,\varphi_{kl}
\end{aligned}$$

The equation shows that F_{ST} can be interpreted as the ratio of the variance of the subpopulation allele frequencies φ_{kl} in relation to the global variance of a Bernoulli distribution. F_{ST} is different for each marker locus and the average F_{ST} is a measure for the distance between the subpopulations.
An important characteristic of this model is that the linkage equilibrium is destroyed in

the total population even if linkage equilibrium between two loci l and l' is assumed in the subpopulations. Some small calculations similar to the above calculations show that the linkage disequilibrium can be calculated as the covariance between the allele frequencies in the same subpopulation at the two different marker loci, i.e.

$$D_{ll'} = \sum_{k=1}^{K} \pi_k(\varphi_{kl} - \varphi_l)(\varphi_{kl'} - \varphi_{l'}) =: \mathrm{Cov}\,(\varphi_{kl}, \varphi_{kl'})$$

Usually, this covariance is not zero, and in the case $K = 2$ it is always unequal to 0 if at both loci the subpopulation allele frequencies differ from each other. Thus, linkage equilibrium is not maintained in the total population.

The subpopulation model with random subpopulation allele frequencies The random subpopulation allele frequency model can be seen as an extension of the previous model which is useful if multiple loci shall be considered in a unique approach. For a fixed locus l, a fixed global allele frequency φ_l is given as in the correlation model. The global allele frequency is interpreted as the allele frequency in a hypothetical ancestral population. The allele frequencies φ_{kl} in the subpopulations are interpreted as iid random variables with $\mathrm{E}\,\varphi_{kl} = \varphi_l$ and $\mathrm{Var}\,\varphi_{kl} = F_k\varphi_l(1-\varphi_l)$. Thus, the model implicitly assumes that the subpopulations all diverged from a common ancestral population at the same time but allows that the subpopulations may have experienced different amount of drift away from the ancestral subpopulation at rates parameterized by F_k. As before we assume random mating and hence Hardy-Weinberg equilibrium in the subpopulations.
Under the assumption that $F_1 = \ldots = F_K = F$ it can be shown that $F = F_{ST}$ as defined in the correlation model. Moreover, the correlation model and the subpopulation model with random allele frequencies are equivalent with respect to expectations and variances. The unconditional expectations and variances in the random subpopulation allele frequency model are the same as in the correlation model. For the expectation of a randomly drawn allele from subpopulation k this can easily be seen from

$$\mathrm{E}\,X_{ilj} = \mathrm{E}\,(\mathrm{E}\,(X_{ilj}|\varphi_{kl})) = \mathrm{E}\,\varphi_{kl} = \varphi_l$$

and for the variance this also can be verified over the conditional variance formula. The covariance for individuals i and i' from the same subpopulation k can be calculated as

$$\begin{aligned}\mathrm{Cov}\,(X_{ilj}, X_{i'lj'}) &= \mathrm{E}\,\mathrm{Cov}\,(X_{ilj}, X_{i'lj'}|\varphi_{kl}, \varphi_{kl}) + \mathrm{Cov}\,(\mathrm{E}\,(X_{ilj}|\varphi_{kl}), \mathrm{E}\,(X_{i'lj'}|\varphi_{kl})) \\ &= 0 + \mathrm{Cov}\,(\varphi_{kl}, \varphi_{kl}) = \mathrm{Var}\,(\varphi_{kl}) = F_k\varphi_l(1-\varphi_l)\end{aligned}$$

and under the assumption $F_1 = \ldots = F_K = F$ it follows that $F = \mathrm{Corr}\,(X_{ilj}, X_{i'lj'}) = F_{ST}$ as defined in the correlation model.
Thus, transforming the variance formula, F_k can be written in the same form as F_{ST} for

the fixed allele frequency model

$$F_k = \frac{\operatorname{Var} \varphi_{kl}}{\varphi_l(1-\varphi_l)} .$$

The equation shows that F_k can be interpreted as the standardized variance of the random variable φ_{kl} which is assumed to be the same for all loci. Thus, as explained before F_k is a measure of the distance of the subpopulation k to the ancestral population. Although the formula has the same form as in the fixed allele frequency model, the interpretation of $\operatorname{Var} \varphi_{kl}$ is different here because this is the specific variance for subpopulation k and not the variance over all subpopulations as in the fixed allele frequency model.
If linkage equilibrium between two loci l and l' is assumed in the subpopulations this theoretically also holds for the total population because

$$\begin{aligned} D = \operatorname{Cov}(X_{ilj}, X_{il'j}) &= \operatorname{E}\operatorname{Cov}(X_{ilj}, X_{il'j}|\varphi_{kl}, \varphi_{kl'}) + \operatorname{Cov}(\operatorname{E}(X_{ilj}|\varphi_{kl}), \operatorname{E}(X_{il'j}|\varphi_{kl'})) \\ &= 0 + \operatorname{Cov}(\varphi_{kl}, \varphi_{kl'}) = 0 \end{aligned}$$

Linkage equilibrium is maintained in the total population because the allele frequencies at the two loci in the same subpopulation are independent from each other in a model with random subpopulation allele frequencies.

The beta-binomial model The beta-binomial model was first applied by Balding and Nichols (1995) for modelling subpopulation allele frequencies. It is a special form of the random subpopulation allele frequency model where it is additionally assumed that the subpopulation allele frequencies φ_{kl} are beta-distributed with distribution function

$$\varphi_{kl} \sim \operatorname{Beta}\left(\frac{1-F_k}{F_k}\varphi_l, \frac{1-F_k}{F_k}(1-\varphi_l)\right).$$

The parameters of the beta distribution are chosen to have the required expectation and variance in the random subpopulation allele frequency model.
If a sample of N_k individuals from subpopulation k is drawn, then the sum of all genotypes $\sum_{i:Z_i=k} X_{il}$ in this sample is Binomial$(2N_k, \varphi_{kl})$-distributed given the subpopulation allele frequency φ_{kl}. However, if φ_{kl} itself is beta-distributed, the unconditional distribution of the sum of all genotypes in this sample is a beta-binomial distribution with

$$\operatorname{E}\left(\sum_{i:Z_i=k} X_{il}\right) = 2N_k\varphi_l, \quad \operatorname{Var}\left(\sum_{i:Z_i=k} X_{il}\right) = (2N_k + 2N_k(2N_k-1)F_k)\varphi_l(1-\varphi_l).$$

2.3.3 Extension of the models to multiple alleles

The concept of inbreeding can easily be extended to a locus l with R_l alleles (Nei, 1977) if $R_l(R_l-1)/2$ inbreeding coefficients $F^{(rr')}$ are defined by a complete specification of the

genotype frequencies dependent on allele frequencies as

$$\begin{aligned} P(B^{(r)}B^{(r)}) &= (\varphi_l^{(r)})^2(1-F^{(rr)}) + \varphi_l^{(r)}F^{(rr)} \\ P(B^{(r)}B^{(r')}) &= 2\varphi_l^{(r)}\varphi_l^{(r')}(1-F^{(rr')}) \quad \text{for} \quad r \leq r'. \end{aligned}$$

To model discrete subpopulations the coefficients $F_{ST}^{(rr')}$, $F_{IT}^{(rr')}$ and $F_{IS}^{(rr')}$ can be defined analogously. In a model with fixed subpopulation allele frequencies there are different methods to summarize these allele-specific inbreeding coefficients into a single measure of inbreeding (Nagylaki, 1998) but it is beyond the scope of this research to explain this here. In a model with random subpopulation allele frequencies it can be assumed that $F_k = F_k^{(rr')}$ is independent of the concrete alleles $B^{(r)}$ and $B^{(r')}$. The beta-binomial model can be extended to the Dirichlet-multinomial model where the vector of allele frequencies $\boldsymbol{\varphi}_{kl}$ is assumed to have a Dirichlet distribution of the form

$$\boldsymbol{\varphi}_{kl} \sim \text{Dirichlet}\left(\frac{1-F_k}{F_k}\varphi_l^{(1)}, \ldots, \frac{1-F_k}{F_k}\varphi_l^{(R_l)}\right).$$

The relationship to the classical definition of Nei (1977) given above can be proven.

2.3.4 Incorporating admixture

Discrete subpopulation models describe a very simple form of population structure. In real populations, however, *admixture* between subpopulations is observed. There is a lot of theory in population genetics which describes the evolution of populations, e.g. described in Gillespie (1998). However, here we only want to consider a model which is not based on the evolutionary theory but on the current structure of the population. In the admixture model each individual is assumed to have inherited some unknown proportion of its alleles from each population. Each individual i is characterized by its admixture proportions which are summarized in a vector $\boldsymbol{q}_i^A = (q_{i1}^A, \ldots, q_{iK}^A)'$ where q_{ik}^A is the proportion of the genome originated from the ancestral subpopulation k for individual i. Thus, the sum of the entries of $\boldsymbol{q}_i^A$ is equal to one. In such a model three sources of linkage disequilibrium can be distinguished if fixed allele frequencies are assumed (Falush et al., 2003). The first source is the *mixture LD* which is caused by variation in the ancestry $\boldsymbol{q}_i^A$ among the sampled individuals. Such variation leads to LD among markers across the genome, even if they are unlinked. This is a generalization of the case of discrete subpopulations where LD can be observed in the total population even if linkage equilibrium in the subpopulations is assumed (section 2.3.2). The second source is the *admixture LD* which additionally occurs between linked markers because individuals are more likely to have alleles from the same subpopulation at linked markers. The explanation is that each chromosome is composed of a set of "chunks" that are derived as an unbroken unit from one of the ancestral populations (see section 2.1). Finally, there is a third source of LD, the *background LD* which occurs between tightly linked markers within subpopulations (see section 2.2.3) and

decays on a much shorter scale (tens of kilobases). However, very complex models are needed to account for each source of linkage disequilibrium and how to infer structure in such models is only briefly discussed in section 5.1.7.

2.4 Inference in subdivided populations

In this section two methods are discussed how the difference between discrete subpopulations can be estimated from a given number of marker loci L in a sample of N individuals if the subpopulation origin of the individuals is known. First a formula for estimation of the fixation index F_{ST} as a classical distance measure is derived and secondly the prediction rate as a further measure of population structure is introduced. Before starting with the details, some vector notation is given.
Altogether, the sample consists of $i = 1, \ldots, N$ individuals. The individuals are genotyped at $l = 1, \ldots, L$ diallelic marker loci. All genotypes can be summarized in a random vector $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_N')'$. The entry $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iL})'$ contains the genotypes of individual i for all marker loci.
The subpopulation allele frequencies are summarized in a vector $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_1', \ldots, \boldsymbol{\varphi}_K')'$ where $\boldsymbol{\varphi}_k = (\varphi_{k1}, \ldots, \varphi_{kL})'$ contains the allele frequencies within subpopulation k for all loci. The vector $\boldsymbol{Z} = (Z_1, \ldots, Z_N)'$ contains for each individual i in the sample the subpopulation where the individual i is from.
Suppose that in each subpopulation k a total of N_k individuals are genotyped at L marker loci $l = 1, \ldots L$. The maximum likelihood estimator for the subpopulation allele frequency φ_{kl} can then be determined as the observed allele frequency in the subpopulation

$$\widehat{\varphi}_{kl} = \frac{1}{2N_k} \sum_{i:Z_i=k} X_{il}.$$

2.4.1 Estimation of the fixation index F_{ST}

In the previous sections it has been shown that the fixation index F_{ST} is an important parameter to describe the distance between subpopulations. In the literature, many methods are proposed to estimate F_{ST} (Nei and Chesser, 1983; Weir and Cockerham, 1984; Weir and Hill, 2002) if genetic marker data are available from individuals of K subpopulations. However, these methods are developed to estimate F_{ST} from data of only one genetic locus and have to be extended in some way if multilocus data are available. Thus, we want to derive here a simple formula to estimate the fixation index F_{ST} between two subpopulations based on the random subpopulation allele frequency model (see section 2.3.2) which can be applied to multilocus marker data. This formula which we have not found to be published elsewhere is applied later in our simulation study (see section 7.1.2).
In the following proposition an unbiased estimator $\widehat{F}_{ST}^*$ for F_{ST} is derived based on the

unrealistic assumption that the ancestral allele frequency φ_l is known. Subsequently, an estimator $\widehat{F}_{ST}$ which can be applied in praxis is given.

Proposition 2.1. *If the ancestral allele frequency φ_l is known, an unbiased estimator for F_{ST} from two subpopulations k and k' is given by*

$$\widehat{F}^*_{ST} = \frac{1}{2 - \frac{1}{2N_k} - \frac{1}{2N_{k'}}} \left(\frac{1}{L} \sum_{l=1}^{L} \frac{(\widehat{\varphi}_{kl} - \widehat{\varphi}_{k'l})^2}{\varphi_l(1-\varphi_l)} - \frac{1}{2N_k} - \frac{1}{2N_{k'}} \right).$$

Proof: First the unconditional expectation and variance of the difference $D_l = \widehat{\varphi}_{kl} - \widehat{\varphi}_{k'l}$ have to be calculated. Since the estimated allele frequencies $\widehat{\varphi}_{kl}$ are unbiased, the unconditional expectation is zero

$$\mathrm{E}\,(D_l) = \mathrm{E}\,(\mathrm{E}\,(D_l|\varphi_{kl}, \varphi_{k'l})) = \mathrm{E}\,(\varphi_{kl} - \varphi_{k'l}) = 0$$

and the unconditional variance can be calculated as

$$\begin{aligned}
\mathrm{Var}\,(D_l) &= \mathrm{E}\,(\mathrm{Var}\,(D_l|\varphi_{kl}, \varphi_{k'l})) + \mathrm{Var}\,(\mathrm{E}\,(D_l|\varphi_{kl}, \varphi_{k'l})) \\
&= \mathrm{E}\,(\mathrm{Var}\,(\widehat{\varphi}_{kl}|\varphi_{kl}) + \mathrm{E}\,(\mathrm{Var}\,(\widehat{\varphi}_{k'l}|\varphi_{k'l}) + \mathrm{Var}\,(\varphi_{kl} - \varphi_{k'l}) \\
&= \frac{1}{2N_k}\mathrm{E}\,\varphi_{kl}(1-\varphi_{kl}) + \frac{1}{2N_{k'}}\mathrm{E}\,\varphi_{k'l}(1-\varphi_{k'l}) + \mathrm{Var}\,(\varphi_{kl}) + \mathrm{Var}\,(\varphi_{k'l}) \\
&= \frac{1}{2N_k}\varphi_l(1-\varphi_l)(1-F) + \frac{1}{2N_{k'}}\varphi_l(1-\varphi_l)(1-F) + 2\varphi_l(1-\varphi_l)F \\
&= \left[\left(\frac{1}{2N_k} + \frac{1}{2N_{k'}} \right)(1-F) + 2F \right] \varphi_l(1-\varphi_l)\,.
\end{aligned}$$

Since $\mathrm{Var}\, D_l = \mathrm{E}\, D_l^2$ it follows

$$\mathrm{E} \left(\frac{1}{L} \sum_{l=1}^{L} \frac{(\widehat{\varphi}_{kl} - \widehat{\varphi}_{k'l})^2}{\varphi_l(1-\varphi_l)} \right) = \left(\frac{1}{2N_k} + \frac{1}{2N_{k'}} \right)(1-F) + 2F$$

and $\mathrm{E}\,\widehat{F}^*_{ST} = F_{ST}$. □

Thus, as an estimator for F_{ST} we propose

$$\widehat{F}_{ST} = \frac{1}{2 - \frac{1}{2N_k} - \frac{1}{2N_{k'}}} \left(\frac{1}{L} \sum_{l=1}^{L} \frac{(\widehat{\varphi}_{kl} - \widehat{\varphi}_{k'l})^2}{\widehat{\varphi}_l(1-\widehat{\varphi}_l)} - \frac{1}{2N_k} - \frac{1}{2N_{k'}} \right)$$

where $\widehat{\varphi}_l$ is the unweighted average of $\widehat{\varphi}_{kl}$ and $\widehat{\varphi}_{k'l}$.

2.4.2 The prediction rate

To understand to what extent it is possible to identify the predefined populations from a given number of marker loci we introduce the *prediction rate* as a new measure for predicting subpopulation membership. We define the prediction rate as the expected posterior

probability of a randomly drawn new individual from one of these samples being correctly classified to its predefined population. If I is a new individual from the predefined (P) population z_I^P the prediction rate can be written as $\mathrm{E}\,(P(Z_I = z_I^P|\boldsymbol{x}_I, \boldsymbol{\varphi}))$. Thus, the prediction rate is an overall measure for the total population considering all subpopulations simultaneously. If there is no difference between the populations a value of 1/K is expected for K populations. If the populations differ from each other the prediction rate should increase with increasing number of marker loci. However, if there is admixture between the different populations the prediction rate does not converge to 100% with a growing number of marker loci. Thus, the prediction rate gives information whether the number of marker loci is sufficiently large to distinguish the subpopulations from each other. In contrast, the classical distance measures like F_{ST} determine the average distance between the subpopulations over the whole genome and the estimated F_{ST} -value does not systematically depend on the number of marker loci. With an increasing number of marker loci only the estimation of F_{ST} becomes more precise. Thus, the prediction rate gives additional information which is not contained in the F_{ST}-value. In section 6.2 the prediction rate is calculated to analyze multilocus marker data from a German study assessing the impact of population stratification within Germany.
The prediction rate can be estimated by leave-one-out crossvalidation. For each individual i the population allele frequencies $\widehat{\boldsymbol{\varphi}}^{(-i)}$ have to be estimated by leaving individual i out. Subsequently, the posterior probability $P(Z_i = z_i^P|\boldsymbol{x}_i, \widehat{\boldsymbol{\varphi}}^{(-i)})$ of an individual i being classified to its predefined population has to be calculated given its multilocus genotype data and the estimated allele frequencies. To determine the posterior probability, Bayes' formula is applied assuming that all subpopulations are a priori equally likely

$$P(Z_i = z_i^P|\boldsymbol{x}_i, \widehat{\boldsymbol{\varphi}}^{(-i)}) = \frac{f(\boldsymbol{x}_i|Z_i = z_i^P, \widehat{\boldsymbol{\varphi}}^{(-i)})}{\sum_{k=1}^{K} f(\boldsymbol{x}_i|Z_i = k, \widehat{\boldsymbol{\varphi}}^{(-i)})}.$$

For each subpopulation k the likelihood for the genotype data of each individual i can be calculated as the product of the corresponding allele frequencies over all loci

$$f(\boldsymbol{x}_i|Z_i = k, \widehat{\boldsymbol{\varphi}}^{(-i)}) = \prod_{l=1}^{L} \binom{2}{x_{il}} (\widehat{\varphi}_{kl}^{(-i)})^{x_{il}} (1 - \widehat{\varphi}_{kl}^{(-i)})^{2-x_{il}}$$

The prediction rate $\mathrm{E}\,(P(Z_I = z_I^P|\boldsymbol{x}_I, \boldsymbol{\varphi}))$ is then estimated by averaging the posterior probabilities over all individuals, i.e.

$$\frac{1}{N} \sum_{i=1}^{N} P(Z_i = z_i^P|\boldsymbol{x}_i, \widehat{\boldsymbol{\varphi}}^{(-i)}).$$

Alternatively, the likelihood could be calculated based on subpopulation genotype frequencies instead of allele frequencies.
Since crossvalidation is performed by subsequently leaving out individuals, a confidence

interval for the prediction rate can be calculated via bootstrapping over all loci. Bootstrap samples of the loci have to be drawn and for the corresponding multilocus data set containing the loci of the bootstrap sample the prediction rate is estimated by the crossvalidation procedure. With the bootstrap procedure the variance in estimating the prediction rate can be calculated and confidence intervals can be derived under the assumption of normality with respect to the bootstrap samples.
To our knowledge we are the first who propose the concept of the prediction rate to describe population differences in the form as introduced here. However, there are similar concepts proposed elsewhere. Paetkau et al. (1997) define the genotype likelihood ratio distance which is also based on the likelihood of the genotype data and estimated applying a crossvalidation procedure. However, it is a distance measure as F_{ST} and not dependent on the number of marker loci as the prediction rate. Thus, the genotype likelihood ratio distance cannot be applied to figure out if the number of marker loci is sufficient. Efron and Tibshirani (1993) describe the general concept of estimating the prediction error via crossvalidation to assess the fit of a statistical model. The prediction error measures the probability of misclassification contrary to the prediction rate. The concept of Efron and Tibshirani (1993) is adopted here for the special context of predicting subpopulation membership from genetic marker data. Usually a disadvantage of crossvalidation is that the variance cannot be easily estimated. In this case the advantage is that not only the individuals but also the loci are independent from each other. Our idea was to additionally implement a bootstrap procedure over the loci.

3 Genetic case-control association studies

This chapter gives an overview about the analysis of genetic case-control association studies. After summarizing the basic principles of case-control studies basic measures for association between a diallelic marker and the disease status as well as some tests for association are described. Furthermore it is discussed how to adjust the analysis if the population is divided into discrete subpopulations and subpopulation membership is known. Finally, it is shown how to apply a logistic regression model to case-control data as an alternative way of analysis. The statistical methods described in this chapter are standard methods for the analysis of contingency tables and are explained in detail for example in Agresti (1996); Collett (2003); Lachin (2000). An overview how these methods can be applied in genetic case-control association studies can be found in Clayton (2000).

3.1 Concepts

3.1.1 General concepts of case-control studies

The basic concepts of case-control studies in epidemiology are summarized to show how the special theory about genetic case-control studies fits into the general epidemiological context. In a *case-control study* a group of affected subjects referred to as cases is compared to a group of unaffected subjects referred to as controls with respect to potential exposures of risk factors. If exposure and disease do not occur independently from each other, an *association* is said to exist between exposure and disease. If there is an association between exposure and disease the distribution of the exposure among cases is different from the distribution of the exposure among controls. From the statistical perspective only associations between exposure and disease can be detected. However, it is of biological interest whether the association between exposure and disease is *causal*, i.e. the exposure has any direct influence on disease development.

A disadvantage of a case-control study is that it is a retrospective study where the exposure is measured after the development of the disease although the true order is different and the subject has to be exposed before developing the disease. However, case-control studies are very important for rare diseases as cancer, for example. Cohort studies generally are too cost- and time consuming in this case. During the follow-up of a cohort of initially healthy subjects only a few cases are expected to occur which is not sufficient for an appropriate statistical analysis unless the cohort is very large.

A critical point is also that cases and controls are usually recruited separately from each other and it may happen that both groups are not comparable with respect to other influence factors. First of all, a systematic *bias* can be introduced when recruiting the study population if cases and controls somehow systematically differ from each other. Such a bias should be avoided by an appropriate design and careful data collection because it

is usually unknown and cannot be corrected for in the statistical analysis. The second problem is that further risk factors may act as *confounders*. A confounder is defined as a variable that leads to an over- or underestimation of the true relation between exposure and disease. The estimate of the effect of the exposure is distorted because it is mixed with the effect of the confounding factor. However, if data on these further risk factors are collected, it is possible to adjust the analysis for potential confounders. There are two necessary criteria for a factor to be a confounder: first of all, a confounder is also a risk factor for the disease and secondly a confounder is associated with the exposure, but not a consequence of the exposure.
Confounding effects have to be controlled for but confounders are not of primary interest as potential risk factors. However, a further type of additional factor occurs if the effect of the exposure on the disease is different for each level of the factor. Such a factor is called *effect modifier* and there is said to be an *interaction* between the exposure and the additional factor. The effect modifier may be a confounder if there is also an association between exposure and effect modifier, but it is not necessarily a confounder.

3.1.2 Concepts of genetic case-control studies

In genetic case-control association studies the exposures are genetic risk factors. Thus, the marker loci which are supposed to be investigated are genotyped for cases and controls. In opposite to the genotype the case-/control-status is often denoted as *phenotype*. If there is an association between the marker and the disease, the distribution of the genotypes is different within cases and controls. There are two different reasons for an association which are of primary interest. The first reason is that the marker locus is itself a disease locus, thus the association is causal. The second reason is that the marker locus is in linkage disequilibrium with the disease locus because both are in close proximity on the genome.
A further question is how marker loci to test for association are chosen. There are basically two approaches of genetic case-control studies with different amount of genotyping, the candidate gene approach and the genome scan. Most of the test statistics we describe here are developed for the candidate gene approach where one or several *candidate genes* are investigated in the study. These candidate genes are often chosen out of biological reasons because they code for some proteins which are expected to be functionally related to the phenotype of interest. On each candidate gene one or several candidate loci are investigated. Thus, we refer to each polymorphism which is genotyped as a separate candidate locus. Most often these candidate loci are single nucleotide polymorphisms and hence diallelic. If several marker loci on the same gene are genotyped they can either be tested separately or jointly for association to the disease. In a joint analysis it has to be accounted for that the marker loci on the same gene usually are in linkage disequilibrium in the population. However, here we only describe tests for single marker analysis.

Although we concentrate here on the candidate gene approach it should be mentioned that in the last years there were rapid technical advances in the development of new genotyping technology making larger investigations feasible. Today SNP arrays are available where for one individual the genotypes of approximately 100,000 single nucleotide polymorphisms can be determined simultaneously. Thus, whole genome association scans are possible. However, it is difficult to interpret the association results from whole genome scans because there is the problem of multiple testing and many false positive association signals are expected to be among the positive association results.
Another source of false positive results either in candidate gene association studies or in genome scans appears due to unobserved confounding. Here, we are especially interested in one special type of confounding caused by population stratification. How to deal with unobserved population stratification in case-control studies is the topic of the following chapters. To motivate the development of methods for unobserved confounding, in this chapter an extension of the basic methods is described to adjust for known confounders in the analysis. Here, we specially concentrate on nominal confounders which can only take few possible values without ordering. In this situation a stratified analysis has to be applied. An example is a population which can be divided into known population strata.

3.2 Analysis of genetic association in a homogeneous population

3.2.1 The genotypic odds ratio and the genotypic relative risks

Given the data in a case-control study, the natural approach is to estimate the genotype distribution given the disease status. For a single diallelic marker with the two alleles B and b the genotype G is uniquely identified by counting the number of B alleles, i.e. $G = 0, 1$ or 2 as described in section 2.2.1. The disease status is described by a random variable Y which takes the value $Y = a$ for the affected population where the cases are sampled from and the value $Y = u$ for the unaffected population where the controls are sampled from. The genotype distribution is described by two vectors $\boldsymbol{f}^{(a)} = (f_0^{(a)}, f_1^{(a)}, f_2^{(a)})'$ and $\boldsymbol{f}^{(u)} = (f_0^{(u)}, f_1^{(u)}, f_2^{(u)})'$ where $f_g^{(y)} = P(G = g|Y = y)$. However, in fact the parameters of interest are the disease risks given the three genotypes which are called *penetrances*. These are denoted by $\phi_g^G = P(Y = a|G = g)$ for $g = 0, 1, 2$. The index G indicates that genotypes are measured as exposure. If allele b is the more common form it would be natural to take genotype bb as a reference category. Thus, two *genotypic relative risks* have to be calculated, i.e.

$$\psi_1^G = \frac{\phi_1^G}{\phi_0^G}, \quad \psi_2^G = \frac{\phi_2^G}{\phi_0^G}.$$

However, a problem in case-control studies is that the penetrances and hence the genotypic relative risks cannot be estimated, because the number of cases and controls is given in advance and the ratio between the number of cases and the total number of individuals in

the study does not correspond to the disease risk in the population. Thus, the measures of association in a case-control study are the *genotypic odds ratios* comparing the odds of disease between the two other genotypes and the reference genotype, i.e.

$$\xi_1^G = \frac{\phi_1^G/(1-\phi_1^G)}{\phi_0^G/(1-\phi_0^G)}, \quad \xi_2^G = \frac{\phi_2^G/(1-\phi_2^G)}{\phi_0^G/(1-\phi_0^G)}.$$

If the disease is rare and all penetrances are small, there is very little difference between the odds ratios and the relative risks and the odds ratio can be used as an approximation for the relative risk.
However, it must be explained why the odds ratio can be estimated in a case-control study. Comparing the odds of disease between a given genotype and the reference is the same as comparing the odds of having that genotype between affecteds and unaffecteds. Thus, the genotypic odds ratios can also be written as

$$\xi_1^G = \frac{f_1^{(a)}/f_0^{(a)}}{f_1^{(u)}/f_0^{(u)}}, \quad \xi_2^G = \frac{f_2^{(a)}/f_0^{(a)}}{f_2^{(u)}/f_0^{(u)}}$$

and can be estimated from genetic case-control data. This is a standard epidemiological result applied in the genetic context. The derivation is based on the relationship between penetrances and genotype distribution within cases and controls. Given the penetrances and additionally the genotype distribution in the total (T) population $\boldsymbol{f}^T = (f_0^T, f_1^T, f_2^T)'$ the genotype distribution within the cases can be calculated via Bayes' formula

$$f_g^{(a)} = \frac{\phi_g^G f_g^T}{\phi}$$

where

$$\phi = \sum_{g'=0}^{2} \phi_{g'}^G f_{g'}^T$$

describes the general disease risk in the population and is commonly denoted as *prevalence*. An analogous formula could be derived for the controls. However, if the disease is rare the genotype distribution in the control group differs little from the genotype distribution in the total population and hence $f_g^T \approx f_g^{(u)}$. This equation can also be inserted in the above formula to calculate $f_g^{(a)}$.

3.2.2 The multiplicative penetrance model and the allelic odds ratio

Under the assumption of Hardy-Weinberg equilibrium in the population (see section 2.2.1) the genotype distribution in the population is completely described by the allelic distribution. Thus, it is sufficient to specify the total allele frequency $p^T = P(B)$. However, even if Hardy-Weinberg equilibrium holds in the total population it is not automatically given within affecteds or unaffecteds if there is an association between the disease and the

genotype. Under the usual assumption of a low disease prevalence, controls should be approximately in HWE, but HWE usually cannot be assumed for cases. Thus, the question arises under which conditions HWE can be assumed for cases and it is sufficient to specify the allele frequencies $p^{(a)} = P(B|Y = a)$ and $p^{(u)} = P(B|Y = u)$ rather than the genotype distribution. Then the *allelic odds ratio* can be used as a measure of association instead of calculating two genotypic odds ratios. The allelic odds ratio is defined as

$$\xi = \frac{p^{(a)}/(1-p^{(a)})}{p^{(u)}/(1-p^{(u)})},$$

analogous to the genotypic odds ratio when written based on genotype frequencies. As proven subsequently, there has to be a multiplicative relationship between the homozygous and heterozygous genotypic relative risks in order to reach HWE within cases. In the so called *multiplicative penetrance model* the homozygous genotypic relative risk is the square of the heterozygous relative risk. Thus, a parameter ψ exists which fulfills the two conditions

$$\psi_1^G = \psi, \quad \psi_2^G = \psi^2$$

or equivalently

$$\phi_1^G = \psi\phi_0^G, \quad \phi_2^G = \psi^2\phi_0^G.$$

This parameter ψ is called *allelic relative risk* because with each additional B-allele the disease risk increases by ψ. The main properties of the multiplicative model are summarized in the following two propositions.

Proposition 3.1. *In the multiplicative penetrance model Hardy-Weinberg equilibrium holds within cases if it can be assumed for the total population. The allele frequency within cases can be calculated as*

$$p^{(a)} = \frac{\psi p^T}{(1-p^T) + \psi p^T}.$$

Proof: Genotype frequencies within cases are calculated under the assumption of HWE in the total population via Bayes' formula as

$$f_0^{(a)} = \frac{\phi_0^G(1-p^T)^2}{\phi}, \quad f_1^{(a)} = \frac{\psi\phi_0^G 2p^T(1-p^T)}{\phi}, \quad f_2^{(a)} = \frac{\psi^2\phi_0^G(p^T)^2}{\phi}$$

where

$$\phi = \phi_0^G(1-p^T)^2 + \psi\phi_0^G 2p^T(1-p^T) + \psi^2\phi_0^G(p^T)^2 = \phi_0^G\left((1-p^T) + \psi p^T\right)^2.$$

Thus, the allele frequency $p^{(a)}$ can be calculated as

$$p^{(a)} = \frac{1}{2}f_1^{(a)} + f_2^{(a)} = \frac{\psi p^T}{(1-p^T) + \psi p^T}.$$

The formulas for HWE within cases can be verified by applying the formulas for genotype- and allele frequencies within cases. □

	$G=0$	$G=1$	$G=2$	
cases	$r_0^{(a)}$	$r_1^{(a)}$	$r_2^{(a)}$	$N^{(a)}$
controls	$r_0^{(u)}$	$r_1^{(u)}$	$r_2^{(u)}$	$N^{(u)}$
total	r_0	r_1	r_2	N

Table 3.1. Observed genotypes for cases and controls

Proposition 3.2. *Assuming a multiplicative penetrance model and Hardy-Weinberg equilibrium in the total population, the allelic relative risk ψ is very closely approximated by the allelic odds ratio ξ if the disease is rare.*

Proof: The odds ratio can be calculated as

$$\xi = \frac{p^{(a)}/(1-p^{(a)})}{p^{(u)}/(1-p^{(u)})} \approx \frac{p^{(a)}/(1-p^{(a)})}{p^T/(1-p^T)} = \frac{\psi p^T/(1-p^T)}{p^T/(1-p^T)} = \psi.$$

□

3.2.3 Commonly applied tests for association

The genotype data of a single diallelic marker with the two alleles B and b can be tabulated in a 2×3 table (table 3.1). Out of total number of N participants r_g individuals have the genotype g. Analogously, the total number of individuals with phenotype y is denoted by $N^{(y)}$ and out of those $r_g^{(y)}$ individuals have genotype g. The null hypothesis of no association in such a table can be formulated in different ways. Since in a case-control study the genotype distribution is estimated conditional on the disease status it is natural to formulate the null hypothesis as $H_0 : \boldsymbol{f}^{(a)} = \boldsymbol{f}^{(u)}$ based on the genotype frequencies. Equivalent formulations are based on the genotypic relative risks $H_0 : \psi_1^G = \psi_2^G = 1$ and genotypic odds ratios $H_0 : \xi_1^G = \xi_2^G = 1$. More generally, the null hypothesis means that genotype and disease are independent, i.e. $H_0 : P(G=g)P(Y=y) = P(G=g, Y=y)$. This is the form of the general null hypothesis for a $r \times c$ table that row and column variable are independent. The null hypothesis can be tested by *Pearson's χ^2-test*. The test statistic is χ^2-distributed with $(r-1)\times(c-1)$ degrees of freedom (df). In this situation Pearson's χ^2-test can be applied with 2 df.
A disadvantage of the χ^2-test is that the power is quite low because of its 2 df. As an alternative *Armitage's trend test* is often applied. This test can be applied for any $2 \times c$ table. Here a score t_c is associated with each column c to detect a special trend in the sequence of odds ratios which are calculated for each column c in comparison to column 1 as a reference. The advantage of the test statistic is that it is χ^2-distributed with only

	allele b	allele B	
cases	$2N^{(a)} - s^{(a)} = 2r_0^{(a)} + r_1^{(a)}$	$s^{(a)} = r_1^{(a)} + 2r_2^{(a)}$	$2N^{(a)}$
controls	$2N^{(u)} - s^{(u)} = 2r_0^{(u)} + r_1^{(u)}$	$s^{(u)} = r_1^{(u)} + 2r_2^{(u)}$	$2N^{(u)}$
total	$2N - s = 2r_0 + r_1$	$s = r_1 + 2r_2$	$2N$

Table 3.2. Observed alleles for cases and controls

1 df. In the genetic context Armitage's trend test is applied with the score $t_g = g$, thus the score simply counts the number of alleles B. This implicitly assumes a multiplicative model of allelic relative risks where an independent effect is associated with each allele.
A third possibility is to apply the allelic χ^2-test. Here the allele distribution is compared between cases and controls. This test statistic is based on the allelic 2×2 table (table 3.2). Here the number of B-alleles is counted for the total sample, denoted by s, as well as for cases and controls separately, denoted by $s^{(y)}, y = a, u$. The null hypothesis $H_0 : p^{(a)} = p^{(u)}$ is tested. This null hypothesis is equivalently formulated as $H_0 : \psi = 1$ or $H_0 : \xi = 1$. The test statistic of the allelic χ^2-test can also be derived from the general formula for Pearson's χ^2-test applied here for a 2×2 table with 1 df. From the results of the last paragraphs it follows that the allelic χ^2-test should only be applied if a multiplicative penetrance model and Hardy-Weinberg equilibrium in the total population can be assumed. A comparison between Armitage's trend test and the allelic χ^2-test shows that both assume a multiplicative penetrance model but Armitage's trend test does not need Hardy-Weinberg equilibrium in the total population. This also becomes clear from a comparison of the test statistics as shown in Sasieni (1997); Devlin and Roeder (1999). Both test statistics have the same numerator which is proportional to the square of the weighted difference between the number of B-alleles in cases and controls $N^{(a)}s^{(u)} - N^{(u)}s^{(a)}$. But the variances are differently calculated. The allelic χ^2-test assumes that all the alleles are independently Bernoulli-distributed which is only true if there is Hardy-Weinberg equilibrium in the population. In contrast, Armitage's trend test accounts for the extra-variance induced by the correlation between the two alleles of one individual. For the derivation of the following methods we want to assume a multiplicative penetrance model and Hardy-Weinberg equilibrium within populations, thus the basic test statistic we apply is the allelic χ^2-test.

3.2.4 Derivation of test statistics for the allelic 2×2 table

As described in the previous section the basic test for the allelic 2×2 table is Pearson's χ^2-test. For a 2×2 table the test statistic can be written in many different ways. Here, we present the formula for the test statistic in the form which is the basis to generalize the test statistic later for stratified analysis. Let $\boldsymbol{G} = (G_1, \ldots, G_N)'$ be the vector of all genotypes

at the candidate locus. The genotype G_i of individual i can be written as $G_i = G_{i1} + G_{i2}$ where G_{ij} is the Bernoulli-distributed random variable which identifies the allele on strand j of individual i as 0 for allele b and 1 for allele B. The number of B-alleles in the table can be written as $s^{(y)} = \sum_{i:Y_i=y} G_i$ and $s = \sum_{i=1}^{N} G_i$ where Y_i is the phenotype of individual i and all phenotypes are summarized in the vector $\boldsymbol{Y} = (Y_1, \ldots, Y_N)'$. The maximum likelihood estimators for the allele frequencies are

$$\widehat{p}^{(y)} = \frac{s^{(y)}}{2N^{(y)}}, \quad \widehat{p} = \frac{s}{2N}$$

separately and together for cases and controls. Note that $\widehat{p}$ is the estimate for the overall allele frequency p under the null hypothesis of no association. The next proposition gives the asymptotic distribution of the test statistic for an allelic 2×2 table. Unless otherwise mentioned, for the asymptotic distribution it is always assumed that $N \to \infty$ and that the ratio of the number of cases and controls is bounded by some constants

$$0 < c^L \leq \frac{N^{(a)}}{N^{(u)}} \leq c^U < \infty.$$

Thus, if $N \to \infty$ it can automatically be concluded that $N^{(a)} \to \infty$ and $N^{(u)} \to \infty$.

Proposition 3.3. *Pearson's χ^2-test statistic for an allelic 2×2 table*

$$\frac{(\widehat{p}^{(a)} - \widehat{p}^{(u)})^2}{\left(\frac{1}{2N^{(a)}} + \frac{1}{2N^{(u)}}\right)\widehat{p}(1-\widehat{p})}$$

is asymptotically χ^2_1-distributed under the null hypothesis of no association $H_0 : p^{(a)} = p^{(u)}$.

Proof: The derivation is based on the fact that $s^{(y)}$ is Binomial$(2N^{(y)}, p^{(y)})$ -distributed under the assumption of HWE. The variance of the estimator $\widehat{p}^{(y)}$ can be directly calculated from the binomial distribution and can be estimated consistently. The standardized allele frequency difference is then asymptotically standard normal distributed. Thus, the square is asymptotically χ^2_1-distributed. □

The allelic χ^2-test can be shown to be a Wald test as described in proposition A.2 with the variance being estimated under the null hypothesis (Lachin, 2000).
It may be argued that the odds ratio is the quantity of interest and hence the test statistic should be based on the estimated odds ratio instead of the allele frequency difference. However, the logarithm of the odds ratio can be written as

$$\log(\widehat{\xi}) = \log\left(\frac{\widehat{p}^{(a)}}{1-\widehat{p}^{(a)}}\right) - \log\left(\frac{\widehat{p}^{(u)}}{1-\widehat{p}^{(u)}}\right).$$

Applying the multivariate delta-method it can be shown that $\sqrt{N}\log(\widehat{\xi})$ is under the null hypothesis asymptotically equally distributed as

$$\sqrt{N}\frac{1}{p(1-p)}(\widehat{p}^{(a)} - \widehat{p}^{(u)})$$

and hence the test statistic based on the log odds ratio derived by Woolf (1955) is asymptotically equivalent to the allelic χ^2-test.
As an alternative to the allelic χ^2-test a likelihood-ratio-test can be applied. The likelihood of the data is calculated as

$$L_1(p^{(a)}, p^{(u)}) = (p^{(a)})^{s^{(a)}}(1-p^{(a)})^{2N^{(a)}-s^{(a)}}(p^{(u)})^{s^{(u)}}(1-p^{(u)})^{2N^{(u)}-s^{(u)}}$$

under the alternative and as

$$L_0(p) = p^s(1-p)^{2N-s}$$

under the null hypothesis.

Proposition 3.4. *The likelihood ratio test statistic*

$$-2\log\left(\frac{L_0(\widehat{p})}{L_1(\widehat{p}^{(a)}, \widehat{p}^{(u)})}\right)$$

is asymptotically χ^2_1-distributed under the null hypothesis of no association $H_0 : p^{(a)} = p^{(u)}$.

Proof: The proposition follows from the asymptotic theory of likelihood ratio tests as described in the appendix A.2.1. □

3.3 Stratified analysis

3.3.1 The bias of the allelic χ^2-test

The simple analysis of allele or genotype tables as introduced before is based on one essential assumption: cases and controls have to be recruited from the same population and must be comparable with respect to other possible risk factors. As already discussed in sections 3.1.1 and 3.1.2 further risk factors may act as confounders and lead to spurious associations between genetic marker and disease if they are not accounted for in the analysis. In this paragraph we want to theoretically show why this happens. We assume that there is a known confounding variable which divides the population into K strata. Here especially the situation is considered where these strata are subpopulations with different genetic structure as described in section 2.3. The distribution of the K subpopulations within cases and controls is given by the following two vectors $\boldsymbol{\pi}^{(a)} = (\pi_1^{(a)}, \ldots, \pi_K^{(a)})'$ and $\boldsymbol{\pi}^{(u)} = (\pi_1^{(u)}, \ldots, \pi_K^{(u)})'$ where $\pi_k^{(y)}$ is the proportion of subpopulation k within individuals of phenotype y, i.e. $\pi_k^{(y)} = N_k^{(y)}/N^{(y)}$ if there are $N_k^{(y)}$ individuals with phenotype y from subpopulation k in the sample. As described in section 2.3.2 there are three different models for discrete subpopulations. In this paragraph we investigate the effect of population stratification in the fixed subpopulation allele frequency model. In this model population stratification results in a bias in the estimators $\widehat{p}^{(a)}$ and $\widehat{p}^{(u)}$ from the allelic χ^2-test. The

null hypothesis is that there is no association in each of the K subpopulations. If the allele frequency distribution within cases and controls is given by $\boldsymbol{p}^{(a)} = (p_1^{(a)}, \ldots, p_K^{(a)})'$ and $\boldsymbol{p}^{(u)} = (p_1^{(u)}, \ldots, p_K^{(u)})'$ and the overall allele frequency distribution by $\boldsymbol{p} = (p_1, \ldots, p_K)'$, the null hypothesis can be formulated as $H_0 : p_k^{(a)} = p_k^{(u)}$ for $k = 1, \ldots, K$. The bias is, for example, calculated in Devlin et al. (2001b) and the formula is given in the next proposition.

Proposition 3.5. *In the fixed subpopulation allele frequency model the expectation of* $\widehat{p}^{(a)} - \widehat{p}^{(u)}$ *is given by*

$$\mathrm{E}\,(\widehat{p}^{(a)} - \widehat{p}^{(u)}) = \sum_{k=1}^{K} (\pi_k^{(a)} - \pi_k^{(u)}) p_k$$

under the null hypothesis of no association.

Proof: The expectation of $\widehat{p}^{(a)} - \widehat{p}^{(u)}$ is calculated as

$$\begin{aligned} E(\widehat{p}^{(a)} - \widehat{p}^{(u)}) &= \frac{1}{2N^{(a)}} \sum_{k=1}^{K} \sum_{i:Z_i=k,Y_i=a} \mathrm{E}\, G_i - \frac{1}{2N^{(u)}} \sum_{k=1}^{K} \sum_{i:Z_i=k,Y_i=u} \mathrm{E}\, G_i \\ &= \sum_{k=1}^{K} \left(\frac{N_k^{(a)}}{N^{(a)}} p_k^{(a)} - \frac{N_k^{(u)}}{N^{(u)}} p_k^{(u)} \right) = \sum_{k=1}^{K} (\pi_k^{(a)} p_k^{(a)} - \pi_k^{(u)} p_k^{(u)}) \end{aligned}$$

and under the null hypothesis the expectation can be simplified to the above formula. □

The bias shows that there are two necessary conditions for population structure to act as a confounder as described in section 3.1.1. The first is that the confounder has to be associated with the disease. Applied to population stratification this condition says that the disease prevalences are different in the subpopulations. This finally leads to a different distribution of the subpopulations between cases and controls. Otherwise, if $\pi_k^{(a)} = \pi_k^{(u)}$ for $k = 1, \ldots K$ the bias is zero. The second condition is that the confounder has to be associated with the exposure. Here, population structure has to be genetically determined at the candidate locus and the allele frequencies at the candidate locus have to be different in the subpopulations.
The effect of the bias is that the allelic χ^2-test has an increased type-I error rate if population stratification exists and is not accounted for. This leads to an increased number of false positive test results if a large number of candidate loci is tested. The next paragraph shows how test statistics could be adjusted for this simple form of population structure as considered here.

3.3.2 Stratified tests for a series of 2×2 tables

If the total population consists of K discrete subpopulations and the subpopulation origin is known for all individuals, the data can be summarized in K allelic 2×2 tables, one for

	allele b	allele B	
cases	$2N_k^{(a)} - s_k^{(a)}$	$s_k^{(a)}$	$2N_k^{(a)}$
controls	$2N_k^{(u)} - s_k^{(u)}$	$s_k^{(u)}$	$2N_k^{(u)}$
total	$2N_k - s_k$	s_k	$2N_k$

Table 3.3. Observed alleles for cases and controls in stratum k.

each subpopulation (table 3.3). The subpopulation allele frequencies can be estimated as before as

$$\widehat{p}_k^{(y)} = \frac{s_k^{(y)}}{2N_k^{(y)}}, \quad \widehat{p}_k = \frac{s_k}{2N_k}$$

separately and pooled for cases and controls. Cochran (1954) generalized the basic form of the χ^2-test statistic as given in proposition 3.3 to the case of a stratified sample. The idea is to use a weighted sum of the estimated subpopulation allele frequency differences in the numerator of the test statistic. The vector $(c_1, \ldots, c_K)'$ contains the weights.

Proposition 3.6. *The Cochran test statistic*

$$\frac{\left(\sum_{k=1}^K c_k(\widehat{p}_k^{(a)} - \widehat{p}_k^{(u)})\right)^2}{\sum_{k=1}^K c_k^2 \left(\frac{1}{2N_k^{(a)}} + \frac{1}{2N_k^{(u)}}\right) \widehat{p}_k(1-\widehat{p}_k)}$$

is asymptotically χ_1^2-distributed for $N_k \to \infty$ under the null hypothesis of no association in the subpopulations.

Proof: The proof simply uses the fact that the sum of independent normal distributed variables is again normal distributed. Since the allele distribution is independent over the subpopulations, the weighted sum of the estimated subpopulation allele frequencies is again asymptotically normal distributed. The square of the standardized weighted sum is then asymptotically χ_1^2-distributed. □

In principle any weights could be taken, but the Cochran test statistic uses the weights

$$c_k = \frac{1}{\frac{1}{2N_k^{(a)}} + \frac{1}{2N_k^{(u)}}}.$$

However, in practice, it is often not the Cochran statistic which is applied for a stratified analysis of a series of 2×2 tables, more popular is the Mantel-Haenszel test statistic (Mantel and Haenszel, 1959). The Mantel-Haenszel test statistic uses a different method for variance estimation. The subpopulation-specific variances are calculated conditional on the observed marginal allele distribution. In this case the number of B-alleles within case-

and control group does not have a binomial distribution but a hypergeometric distribution under the null hypothesis of no association. However, although the Mantel-Haenszel test statistic is more popular, the unconditional approach has the advantage that it can be applied more generally. This is for example discussed in Miettienen (1985). Anyway, asymptotically both tests are equivalent and are often referred to interchangeably as Cochran-Mantel-Haenszel test.
It can be shown that the Cochran-Mantel-Haenszel test is designed to test the null hypothesis

$$H_0 : \xi_1 = \ldots = \xi_K = 1$$

against a restricted alternative hypothesis of a common odds ratio unequal to 1 in all subpopulations

$$H_1 : \xi_1 = \ldots = \xi_K = \xi^{(0)}, \quad \xi^{(0)} \neq 1$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)'$ is the vector of the allelic odds ratios in all subpopulations. Thus, the test statistic may have low power for other alternatives. How such a restricted alternative hypothesis is interpreted is discussed at the end of the section. Here, we want to give an idea why the test statistic is constructed for the restricted alternative but do not want to explain this formally. First of all the alternative hypothesis can be written in the form

$$H_1 : g(p_k^{(a)}) - g(p_k^{(u)}) = \delta, \quad \delta \neq 0, \quad \text{for} \quad k = 1, \ldots, K$$

where $g(p) = \log(\frac{p}{1-p})$ and $\delta = \log(\xi^{(0)})$ and the null hypothesis can be formulated analogously with $\delta = 0$. The numerator of the test statistic

$$W = \sum_{k=1}^{K} c_k(\widehat{p}_k^{(a)} - \widehat{p}_k^{(u)})$$

is an unbiased estimator for the common log odds ratio δ multiplied by a constant. This can be shown by applying the intermediate value theorem

$$E(W) = \sum_{k=1}^{K} c_k(p_k^{(a)} - p_k^{(u)}) = \sum_{k=1}^{K} \frac{c_k(g(p_k^{(a)}) - g(p_k^{(u)}))}{g'(p_k^*)} = \delta \sum_{k=1}^{K} \frac{c_k}{g'(p_k^*)}$$

where g' is the derivative of g and p_k^* is a value lying between $p_k^{(a)}$ and $p_k^{(u)}$. Furthermore it can be shown that the weights c_k of the Cochran-Mantel-Haenszel test statistic are chosen to maximize the Pitman efficiency (Noether, 1955) of the test statistic for the alternative considered here (Radhakrishna, 1965). Roughly formulated, the Cochran-Mantel-Haenszel test statistic is asymptotically most powerful among all these weighted test statistics considering a limiting process where the alternative hypothesis converges against the null hypothesis.
To test the null hypothesis

$$H_0 : \boldsymbol{p}^{(a)} = \boldsymbol{p}^{(u)} \quad \text{or equivalently} \quad H_0 : \boldsymbol{\xi} = 1$$

against the unrestricted alternative

$$H_1 : \boldsymbol{p}^{(a)} \neq \boldsymbol{p}^{(u)} \quad \text{or equivalently} \quad H_1 : \boldsymbol{\xi} \neq \mathbf{1}$$

Pearson's χ^2-test statistics for all 2×2 tables can be summed up over the subpopulations. The main difference of this test statistic to the Cochran-Mantel-Haenszel test is that in this case the allele frequency differences are squared and standardized before calculating the sum.

Proposition 3.7. *The test statistic*

$$\sum_{k=1}^{K} \frac{(\widehat{p}_k^{(a)} - \widehat{p}_k^{(u)})^2}{\left(\frac{1}{2N_k^{(a)}} + \frac{1}{2N_k^{(u)}}\right) \widehat{p_k}(1 - \widehat{p_k})}$$

is asymptotically χ^2_K-distributed for $N_k \to \infty$ under the null hypothesis of no association in the subpopulations.

Proof: Since the alleles in the different subpopulations are independent from each other, the test statistic is a sum of independent χ^2_1-distributed test statistics and hence asymptotically χ^2_K-distributed. □

As an extension of Pearson's χ^2-test to a stratified sample the Cochran-Mantel-Haenszel test as well as the corresponding test for the unrestricted alternative can analogously shown to be Wald tests with the variance being estimated under the null hypothesis.
For a stratified sample also a likelihood ratio test can be applied. The likelihood of the genotype data at the candidate locus is calculated as

$$L_1(\boldsymbol{p}^{(a)}, \boldsymbol{p}^{(u)}) = \prod_{k=1}^{K} (p_k^{(a)})^{s_k^{(a)}} (1 - p_k^{(a)})^{2N_k^{(a)} - s_k^{(a)}} (p_k^{(u)})^{s_k^{(u)}} (1 - p_k^{(u)})^{2N_k^{(u)} - s_k^{(u)}}$$

under the alternative and as

$$L_0(\boldsymbol{p}) = \prod_{k=1}^{K} p_k^{s_k} (1 - p_k)^{2N_k - s_k}$$

under the null hypothesis. There are two versions of a likelihood ratio test, one for the restricted and the other for the unrestricted alternative. The conditions how to determine the maximum likelihood estimator under the alternative differ from each other in both cases. The likelihood can be maximized either restrictedly or unrestrictedly. If the maximum likelihood estimator is calculated under the restriction that the estimated odds ratio is the same in all subpopulations the likelihood function is a function of the allele frequencies within controls $\boldsymbol{p}^{(u)}$ and the common odds ratio $\xi^{(0)}$. The maximum likelihood

estimators than have to be determined iteratively. Restricted maximum likelihood estimation is for example proposed by Miettienen (1985). As for the Wald tests unrestricted and restricted maximum likelihood estimation lead to a different asymptotic distribution of the test statistic.

Proposition 3.8. *The unrestricted likelihood ratio test statistic*

$$-2\log\left(\frac{L_0(\widehat{\boldsymbol{p}})}{L_1(\widehat{\boldsymbol{p}}^{(a)},\widehat{\boldsymbol{p}}^{(u)})}\right)$$

is asymptotically χ^2_K-distributed under the null hypothesis $H_0 : \boldsymbol{p}^{(a)} = \boldsymbol{p}^{(u)}$.
The restricted likelihood ratio test statistic

$$-2\log\left(\frac{L_0(\widehat{\boldsymbol{p}})}{L_1(\widehat{\boldsymbol{p}}^{(u)},\widehat{\xi}^{(0)})}\right)$$

is asymptotically χ^2_1-distributed under the null hypothesis $H_0 : \boldsymbol{p}^{(a)} = \boldsymbol{p}^{(u)}$.

Proof: The distribution of the test statistics follows from the asymptotic theory of likelihood ratio tests as described in the appendix A.2.1. Most important is the difference in the degrees of freedom in the test statistic. In the case of restricted maximum likelihood estimation only one additional parameter has to be estimated under the alternative compared to the null hypothesis whereas in the first case there are K additional parameters. □

The choice of the alternative depends on the assumptions of the type of population stratification. The question is if population stratification only acts as a confounder or if population structure also could be an effect modifier if a genetic effect exists. In the first case the genetic effect is assumed to be the same in all subpopulations whereas in the latter case the candidate genes show a different effect in each subpopulation and hence the penetrances and the odds ratios differ in the subpopulations. In other words, in the first case *homogeneity* of genetic effects in the subpopulations is assumed whereas in the second case *heterogeneity* between the subpopulations is allowed. Hence, if only a confounding effect is expected to occur the test statistics for the restricted alternative (Cochran-Mantel-Haenszel test or the respective likelihood ratio test) should be employed because of their higher power for the restricted alternatives. However, if effect modification is possible the test statistics for the unrestricted alternative are the first choice. Of course, it also could be possible that there is no confounding effect at all and only an effect modification is observed. However, such a situation is not as problematic as a confounding effect because in this situation no excess of false positives is expected to occur under the null hypothesis if population structure is not accounted for. Thus, we concentrate here on the confounding effect and a possible additional effect modification. In which cases such an effect modification could exist is a biological question which is not fully answered up to date.

However, for more subtle population structure as considered here the effect modification may be small if at all present and thus we recommend testing the null hypothesis against the restricted alternative. Out of the two test statistics for the restricted alternative we would prefer the Cochran-Mantel-Haenszel test in comparison to the restricted likelihood ratio test. The Cochran-Mantel-Haenszel test should be less sensitive if this assumption is not valid because the subpopulation allele frequencies in case and control group are not estimated under the restricted alternative.

3.4 Logistic regression

3.4.1 The logistic regression model for case-control data

An alternative way for the analysis of case-control data is to apply logistic regression. In a logistic regression model applied for epidemiological data it is natural to model the probability of developing disease given the exposure. However, as already explained in section 3.2.1 this probability cannot be estimated in a case-control study because the number of cases and controls is fixed in advance. Thus, in a case-control study the logistic regression model is based on the probability of an individual of being a case in the case-control sample. An advantage of the logistic regression model is that many exposures can be incorporated in the model. Hence, genetic and non-genetic factors can be analyzed simultaneously. Let $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iQ})'$ be the vector of the Q exposures of individual i. Then the logit of probability $\phi_i = P(Y_i = a|\boldsymbol{v}_i)$ of individual i being a case in the case-control sample is modelled by a linear function of the exposures $\boldsymbol{v}_i$. The logistic regression has the form

$$\operatorname{logit} \phi_i = \alpha + \boldsymbol{v}_i' \boldsymbol{\beta}$$

where the logit function is defined as

$$\operatorname{logit} \phi_i = \log \left(\frac{\phi_i}{1 - \phi_i} \right)$$

and the intercept α and the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_Q)'$ have to be estimated. As already explained, the probability ϕ_i is modelled but the probability ϕ_i^T of developing the disease in the total population is the quantity of interest. It can be shown (e.g. Breslow and Day, 1980) that only the intercept α in the regression model changes if the model is applied for ϕ_i instead of ϕ_i^T and the regression coefficients $\boldsymbol{\beta}$ remain the same. Based on that argument logistic regression can be applied for case-control data.
The likelihood for the logistic regression model is the likelihood for the phenotype data. To formulate the likelihood an indicator variable Y_i^I for the phenotype has to be defined, $Y_i^I = 0$ for the controls and $Y_i^I = 1$ for the cases. The likelihood then has the form

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{N} f(y_i^I | \boldsymbol{v}_i) = \prod_{i=1}^{N} \phi_i^{y_i^I} (1 - \phi_i)^{1 - y_i^I}$$

where the probabilities ϕ_i can be obtained by the logistic function (inverse logit)

$$\phi_i = \frac{\exp(\alpha + \boldsymbol{v}_i'\boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{v}_i'\boldsymbol{\beta})}.$$

To determine the maximum likelihood estimates for α and $\boldsymbol{\beta}$ an iterative procedure such as the Newton-Raphson algorithm has to be applied because a system of non-linear equations has to be solved. Wald tests, likelihood ratio tests and score tests can be derived in the model, for the global null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ of no association with any of the exposures as well as for the null hypothesis $H_0 : \beta_q = 0$ of no association with exposure q.

The exposures can be qualitative or quantitative. Binary exposures are coded with the numbers 0 or 1, i.e. for a binary exposure q the variable v_{iq} is equal to 0 or 1. Furthermore a standard epidemiological result is (Lachin, 2000) that the odds ratio of developing the disease between exposed and non-exposed individuals is $\xi_q = \exp(\beta_q)$. If the exposures are nominal, a dummy coding has to be introduced. For a nominal exposure with T levels denoted as $1, \ldots, T$, for each individual a vector of $T-1$ binary variables $\boldsymbol{d}_i = (d_{i1}, \ldots d_{i,T-1})'$ has to be introduced. One level has to be defined as a reference, for example level 1. Then $d_{i,t-1} = 1$, if individual i has the exposure t for $t = 2, \ldots, T$ and $d_{it} = 0$ otherwise. Thus, for the reference level all of the dummy variables are zero and for all the other levels always one of the dummy variables is equal to 1. Odds ratios between all other categories and the reference category can be calculated.

3.4.2 The logistic regression model applied to genetic data

If a diallelic genetic marker is considered as an exposure and no further exposures are included in the model, there are two possibilities to include this marker into the model, either with dummy coding or without dummy coding. If dummy coding is applied the genotype g_i, originally coded as $0, 1, 2$ is then contained in a dummy vector $\boldsymbol{g}_i^G = (g_{i1}^G, g_{i2}^G)'$ of two binary variables as described before, where 0 is the reference category. The model takes the form

$$\operatorname{logit} \phi_i = \alpha^G + (\boldsymbol{g}_i^G)'\boldsymbol{\beta}^G$$

where $\boldsymbol{\beta}^G = (\beta_1^G, \beta_2^G)'$ and $\xi_1^G = \exp(\beta_1^G)$ and $\xi_2^G = \exp(\beta_2^G)$ are the genotypic odds ratios of the other two genotypes compared to the reference genotype. Under the assumption of a multiplicative model of odds ratios the original coding can be directly applied and the model can be written as

$$\operatorname{logit} \phi_i = \alpha + g_i\beta.$$

The odds ratio $\xi = \exp(\beta)$ is equal to the heterozygous odds ratio and the homozygous odds ratio is the square of the heterozygous odds ratio, i.e.

$$\xi_1^G = \xi, \quad \xi_2^G = \xi^2.$$

If the disease is rare, a model of multiplicative odds ratios is very close to a multiplicative penetrance model as described in section 3.2.2. Since logistic regression is always based on genotypes instead of alleles, Hardy-Weinberg equilibrium is not assumed. If the original coding is applied, the score test for $H_0 : \beta = 0$ is equivalent to Armitage's trend test described in section 3.2.3.
If the population is not homogeneous and consists of several subpopulations, the subpopulation an individual is from has to be included in the model as an additional nominal exposure. If the parameters for the subpopulations are contained in the vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)'$ with the constraint that $\sum_{k=1}^{K} \eta_k = 0$ the logistic regression model based on original genotypes has the form

$$\operatorname{logit} \phi_i = \alpha + g_i\beta + \eta_{z_i}.$$

To formulate the logistic regression in the general form as introduced above, for each individual i a vector of $K-1$ binary dummy variables $\boldsymbol{z}_i^D = (z_{i1}^D, \ldots z_{i,K-1}^D)'$ has to be defined which codes the subpopulation z_i. Then the parameters for the subpopulations are contained in a $K-1$-dimensional vector $\boldsymbol{\eta}^D = (\eta_1^D, \ldots, \eta_{K-1}^D)'$ and the model can be written as

$$\operatorname{logit} \phi_i = \alpha + g_i\beta + (\boldsymbol{z}_i^D)'\boldsymbol{\eta}^D.$$

If the subpopulations are only included as an additional term in the model it is implicitly assumed that population stratification only acts as a confounder and not as an effect modifier. Effect modifications could also be included in the logistic regression by considering an additional term in the model based on the multiplication of exposures. Thus, an additional $K-1$-dimensional parameter vector has to be estimated modelling the difference in the genetic effect in each subpopulation in comparison to the reference subpopulation. However, such an interaction term is usually omitted because this is expected to lead to a loss of power comparable to classical tests for the unrestricted alternative. Moreover, the meaning of the additional interaction parameters has to be carefully interpreted in the logistic model.

4 Genomic Control

The first approach proposed in the literature to test for association in a case-control study adjusting for unobserved population stratification is the method of Genomic Control (GC). The alternative concept of Structured Association (SA) is discussed in the next chapter. Both concepts have in common that a set of additional marker loci has to be genotyped to estimate the amount of population structure in the sample and to correct the tests for association in an appropriate way. However, both concepts make different use of these additionally genotyped null loci. The idea of Genomic Control is to apply the usual test statistics for the test of association but to estimate the variance empirically from the additional multilocus marker data to correct for unobserved population stratification. Genomic Control was originally proposed by Devlin and Roeder (1999). Later on different versions of Genomic Control were discussed (Reich and Goldstein, 2001; Devlin et al., 2004) and its power and properties were investigated by several authors theoretically and in simulations (Bacanu et al., 2000; Devlin et al., 2001a,b; Marchini et al., 2004; Shmulewitz et al., 2004). The concept of Genomic Control is based on the correlation model for general population structure as described in section 2.3.1. Before introducing the method, we would like to investigate the effect of population stratification on the allelic χ^2-test in the correlation model.

4.1 The variance inflation of the allelic χ^2-test

As described in section 2.3.1 in the correlation model a global allele frequency $p^{(a)}$ for cases and a global allele frequency $p^{(u)}$ for controls is modelled for the candidate locus. Similarities between individuals are modelled over correlations between alleles from closely related individuals. Correlations within individuals are modelled over the inbreeding coefficients F_i and correlation between individuals over the kinship coefficients $f_{ii'}$. Correlations are allowed to be different for all individuals. If the difference $\widehat{p}^{(a)} - \widehat{p}^{(u)}$ from the numerator of the allelic χ^2-test is taken, the expectation under the null hypothesis of no association remains zero but the variance is inflated. This has been proven by Devlin et al. (2001b) for equal sample sizes of the case and control group. Here we extend their calculation to different sample sizes of cases and controls. The inflation factor is given relative to the variance which is estimated in the allelic χ^2-test.

Proposition 4.1. *In the correlation model for general population structure the variance of $\widehat{p}^{(a)} - \widehat{p}^{(u)}$ is inflated by the variance inflation factor*

$$\lambda = 1 - \frac{\frac{1}{2N^{(a)}}\overline{F}^{(a)} + \frac{1}{2N^{(u)}}\overline{F}^{(u)}}{\frac{1}{2N^{(a)}} + \frac{1}{2N^{(u)}}} + \frac{1}{\frac{1}{2N^{(a)}} + \frac{1}{2N^{(u)}}}\left(\overline{f}^{(a)} + \overline{f}^{(u)} - 2\overline{f}^{(au)}\right)$$

under the null hypothesis of no association where

$$\begin{aligned}\overline{F}^{(y)} &= \frac{1}{N^{(y)}}\sum_{i:Y_i=y} F_i, \quad \overline{f}^{(y)} = \frac{1}{(N^{(y)})^2}\sum_{i:Y_i=y}\left(F_i + \sum_{i'\neq i:Y_{i'}=y} f_{ii'}\right) \quad \textit{for} \quad y = a, u,\\ \overline{f}^{(au)} &= \frac{1}{N^{(a)}N^{(u)}}\sum_{i:Y_i=a}\sum_{i':Y_{i'}=u} f_{ii'}.\end{aligned}$$

Proof: First of all the variance within one individual can be calculated as

$$\begin{aligned}\mathrm{Var}\,(G_i) &= \mathrm{Var}\,(G_{i1}) + \mathrm{Var}\,(G_{i2}) + 2\mathrm{Cov}\,(G_{i1}, G_{i2})\\ &= 2p(1-p) + 2F_i p(1-p) = 2(1+F_i)p(1-p)\end{aligned}$$

and the covariance between two individuals is given by

$$\begin{aligned}\mathrm{Cov}\,(G_i, G_{i'}) &= \mathrm{Cov}\,(G_{i1}, G_{i'1}) + \mathrm{Cov}\,(G_{i1}, G_{i'2}) + \mathrm{Cov}\,(G_{i2}, G_{i'1}) + \mathrm{Cov}\,(G_{i2}, G_{i'2})\\ &= 4f_{ii'}p(1-p).\end{aligned}$$

For the variance of the observed allele frequency difference then follows

$$\begin{aligned}\mathrm{Var}\,(\widehat{p}^{(a)} - \widehat{p}^{(u)}) &= \mathrm{Var}\left(\frac{1}{2N^{(a)}}\sum_{i:Y_i=a} G_i - \frac{1}{2N^{(u)}}\sum_{i:Y_i=u} G_i\right)\\ &= \Bigg[\frac{1}{4(N^{(a)})^2}\sum_{i:Y_i=a} 2(1+F_i) + \frac{1}{4(N^{(u)})^2}\sum_{i:Y_i=u} 2(1+F_i)\\ &+ \frac{1}{4(N^{(a)})^2}\sum_{i:Y_i=a}\sum_{i'\neq i:Y_{i'}=a} 4f_{ii'} + \frac{1}{4(N^{(u)})^2}\sum_{i:Y_i=u}\sum_{i'\neq i:Y_{i'}=u} 4f_{ii'}\\ &- 2\frac{1}{4N^{(a)}N^{(u)}}\sum_{i:Y_i=a}\sum_{i':Y_{i'}=u} 4f_{ii'}\Bigg] p(1-p)\\ &= \left[\frac{1}{2N^{(a)}} + \frac{1}{2N^{(u)}} - \frac{1}{2N^{(a)}}\overline{F}^{(a)} - \frac{1}{2N^{(u)}}\overline{F}^{(u)} + \overline{f}^{(a)} + \overline{f}^{(u)} - 2\overline{f}^{(au)}\right] p(1-p)\end{aligned}$$

The variance inflation is calculated with respect to the variance in a homogeneous population

$$p(1-p)\left(\frac{1}{2N^{(a)}} + \frac{1}{2N^{(u)}}\right).$$

□

The formula for the variance inflation can be simplified in the special case of discrete subpopulations where the same correlation F_{IT} within all individuals as well as the same correlation F_{ST} between all individuals in the same subpopulation is assumed. As described in section 2.3.2 we assume HWE in the subpopulations and hence $F_{ST} = F_{IT}$.

Proposition 4.2. *In the correlation model for discrete subpopulations the variance of* $\widehat{p}^{(a)} - \widehat{p}^{(u)}$ *is inflated by the variance inflation factor*

$$\lambda = 1 - F_{ST} + \frac{1}{\frac{1}{2N^{(a)}} + \frac{1}{2N^{(u)}}} F_{ST} \sum_{k=1}^{K} (\pi_k^{(a)} - \pi_k^{(u)})^2$$

under the null hypothesis of no association.

Proof: In this special case the average correlations are

$$\begin{aligned} \overline{F}^{(y)} &= F_{ST}, \quad \overline{f}^{(y)} = \frac{1}{(N^{(y)})^2} \sum_{k=1}^{K} (N_k^{(y)})^2 F_{ST} = \sum_{k=1}^{K} (\pi_k^{(y)})^2 F_{ST} \quad \text{for} \quad y = a, u \\ \overline{f}^{(au)} &= \frac{1}{N^{(a)} N^{(u)}} \sum_{k=1}^{K} N_k^{(a)} N_k^{(u)} F_{ST} = \sum_{k=1}^{K} \pi_k^{(a)} \pi_k^{(u)} F_{ST} \end{aligned}$$

and for the variance it follows

$$\mathrm{Var}\,(\widehat{p}^{(a)} - \widehat{p}^{(u)}) = p(1-p) \left[\left(\frac{1}{2N^{(a)}} + \frac{1}{2N^{(u)}} \right) (1 - F_{ST}) + F_{ST} \sum_{k=1}^{K} (\pi_k^{(a)} - \pi_k^{(u)})^2 \right].$$

The variance inflation is again calculated with respect to the variance in a homogeneous population. □

It should be mentioned that the same variance inflation can be derived in the random subpopulation allele frequency model since we proved in section 2.3.2 that both models are equivalent with respect to expectations and variances. Thus, in this model the effect of population stratification is also a variance inflation instead of a bias.
The formula for the variance inflation also shows that the distribution of the subpopulations has to be different between cases and controls to have a possible confounding effect. Otherwise, if $\pi_k^{(a)} - \pi_k^{(u)} = 0$ for $k = 1, \ldots K$ the variance inflation in proposition 4.2 is even slightly smaller than 1. The second condition for confounding is that the population structure has to be genetically determined, i.e. F_{ST} has to be unequal to zero.
It is also important to point out the impact of the variance inflation for large case-control studies. The variance of $\widehat{p}^{(a)} - \widehat{p}^{(u)}$, calculated in the proof of proposition 4.2, does not converge against zero for $N \to \infty$. Instead, there is always the fixed variance term

$$p(1-p) F_{ST} \sum_{k=1}^{K} (\pi_k^{(a)} - \pi_k^{(u)})^2$$

left. This term depends on F_{ST} and on the sum of the squared differences of the subpopulation proportions within cases and controls. This sum lies between 0 and 2 and the maximum is reached if cases and controls form two different subpopulations themselves, then $\pi_1^{(a)} = 1, \pi_2^{(a)} = 0$ and $\pi_1^{(u)} = 0, \pi_2^{(u)} = 1$. Dependent on the amount of population

structure case-control studies could reach sample sizes where the fixed variance term is large in comparison to the sample-size dependent term

$$p(1-p)\left(\frac{1}{2N^{(a)}}+\frac{1}{2N^{(u)}}\right)(1-F_{ST}).$$

For these large sample sizes the variance reduction of the test statistic is very small with additional recruitment of individuals. Thus, an additional recruitment does not make sense anymore if there are no other methods available to adjust for population stratification in the analysis.

The formula for the variance inflation factor says that $\lambda-(1-F_{ST})$ is proportional to the harmonic mean of $N^{(a)}$ and $N^{(u)}$ and proportional to F_{ST}. Again, it becomes visible that the problem of variance inflation is more relevant in larger case-control studies which are necessary to detect small genetic effects. Furthermore the variance inflation is dependent on the number of subpopulations and is smaller if a population consists of a large number of subpopulations.

However, it must be mentioned that the distribution theory for a homogeneous population cannot be generally transferred to a structured population. In a homogeneous population the allelic χ^2-test is based on the asymptotic standard normal distribution of

$$\frac{\widehat{p}^{(a)}-\widehat{p}^{(u)}}{\sqrt{\left(\frac{1}{2N^{(a)}}+\frac{1}{2N^{(u)}}\right)p(1-p)}}$$

but in a structured population

$$\frac{\widehat{p}^{(a)}-\widehat{p}^{(u)}}{\sqrt{\lambda\left(\frac{1}{2N^{(a)}}+\frac{1}{2N^{(u)}}\right)p(1-p)}}$$

is generally not asymptotically standard normal distributed because the allele frequencies $\widehat{p}^{(a)}$ and $\widehat{p}^{(u)}$ are not the average of independent alleles anymore. However for the special case of a discrete subpopulation model and a growing number of subpopulations K the asymptotic normal distribution can be proven. Similar considerations can be found in Devlin et al. (2001b) but the underlying subpopulation model is different.

Proposition 4.3. *In the correlation model for discrete subpopulations*

$$\frac{\widehat{p}^{(a)}-\widehat{p}^{(u)}}{\sqrt{\lambda\left(\frac{1}{2N^{(a)}}+\frac{1}{2N^{(u)}}\right)p(1-p)}}$$

is standard normal distributed under the null hypothesis of no association for $K\longrightarrow\infty$, $N_k\leq c$ for $k=1,\dots K$ and a constant $c<\infty$.

Proof: The allele frequency difference multiplied by $2N$ can be written as

$$2N(\widehat{p}^{(a)} - \widehat{p}^{(u)}) = \sum_{k=1}^{K} T_k$$

where

$$T_k = \frac{N}{N^{(a)}} \sum_{i:Y_i=a,Z_i=k} X_i - \frac{N}{N^{(u)}} \sum_{i:Y_i=u,Z_i=k} X_i.$$

The variables T_k are independent from each other and bounded since $N_k < c$. The variance

$$\mathrm{Var}\left(2N(\widehat{p}^{(a)} - \widehat{p}^{(u)})\right) = p(1-p)\left[2N\left(\frac{N}{N^{(a)}} + \frac{N}{N^{(u)}}\right)(1-F) + 4N^2F\sum_{k=1}^{K}(\pi_k^{(a)} - \pi_k^{(u)})^2\right]$$

as derived in proposition 4.2 converges against infinity for $K \longrightarrow \infty$ and hence $N \longrightarrow \infty$ because of the first term. Thus, the asymptotic normal distribution follows from a corollary of the Lindeberg-Feller central limit theorem (Karr, 1993) because the Lindeberg condition is satisfied for a uniformly bounded sequence of random variables. □

4.2 The method of Genomic Control

4.2.1 The general concept

As shown in the previous section the effect of population stratification is that the variances of the usual test statistics are inflated. We only considered the allelic χ^2-test there, but the variance inflation could similarly be calculated relative to the variance which is estimated in Armitage's trend test. Thus, if population stratification exists, an *overdispersion* of the test statistics can be observed. The idea of Genomic Control is to use the additionally genotyped marker loci to empirically estimate the variance inflation under the null hypothesis of no association. In the original version of Genomic Control Devlin and Roeder (1999) proposed to use Armitage's trend test statistic as test statistic since it automatically accounts for the extra-variance which occurs if the total population is not in Hardy-Weinberg equilibrium (section 3.2.3). Reich and Goldstein (2001) considered the allelic χ^2-test statistic instead. In both cases the test statistic, denoted by T^2, is asymptotically χ^2_1-distributed if no population stratification exists. To account for the variance inflation Devlin and Roeder (1999) suggested to approximate the distribution of the test statistic under the null hypothesis by a scaled χ^2_1-distribution. The scaling factor is the variance inflation λ which has to be estimated. Hence they assume that only the variance is inflated but the shape of the distribution does not change. As discussed in the previous section this is only an approximation because the asymptotic standard normal distribution of the standardized allele frequency difference is theoretically not maintained if population stratification exists. However, Devlin and Roeder (1999) nevertheless propose to use the

standard normal distribution as an approximation for the distribution of $T/\sqrt{\lambda}$ or equivalently the χ^2_1-distribution as an approximation for the distribution of T^2/λ. For some forms of population structure the approximation may break down in the extreme tails of the distribution but this has not been described so far as a major problem. Reich and Goldstein (2001) empirically investigate the distribution of the allelic χ^2-test statistic in the presence of stratification by simulating two subpopulations and varying the distance between the subpopulations. Their result is that the scaled χ^2_1-distribution fits quite well unless in the case of extreme stratification. Thus, we assume that the test statistic is approximately scaled χ^2_1-distributed for the following derivations.

The remaining question is how to estimate the variance inflation factor λ. The idea is to calculate the same test statistics $T_1^2, \ldots, T_L^2$ as for the candidate locus also for the additionally genotyped null loci $l = 1, \ldots, L$. Then the inflation is estimated in comparison to the theoretical value in a homogeneous population. Devlin and Roeder (1999) proposed to take the median of $T_1^2, \ldots, T_L^2$ divided by the median of the χ^2_1-distribution to estimate λ. Reich and Goldstein (2001) suggested to take the mean-based estimator $\widehat{\lambda} = \frac{1}{L}\sum_{l=1}^{L} T_l^2$ instead. The mean does not have to be corrected by a factor because the χ^2_1-distribution has the expectation 1. The estimator $\widehat{\lambda}$ is changed to 1 if a value below 1 is estimated because values below 1 are very unlikely. However, there are situations where the inflation factor λ could be slightly smaller than 1 (see proposition 4.2), but values clearly smaller than 1 are not reasonable. If the mean is taken to estimate the variance inflation factor the advantage is that the distribution of the estimator can be easily determined. Since the sum of L independent χ^2_1-distributed random variables is χ^2_L-distributed, it follows for the mean-based estimator that $\widehat{\lambda}$ is approximately $\frac{\lambda}{L}\chi^2_L$-distributed. This formula can be applied to derive a confidence interval for λ, as shown below. Furthermore, even without explicitly defining the inflation factor the statistic $\sum_{l=1}^{L} T_l^2$ can be used as test statistic for the null hypothesis that there is no confounding effect by population stratification (Pritchard and Rosenberg, 1999). Under the null hypothesis of no variance inflation it follows that $\sum_{l=1}^{L} T_l^2$ is asymptotically χ^2_L-distributed.

To eventually test for association in the presence of population stratification the Genomic Control test statistic is applied dividing the original test statistic by the estimated variance inflation. The test statistic $T^2/\widehat{\lambda}$ is assumed to be approximately χ^2_1-distributed. Devlin and Roeder (1999) used Armitage's trend test as original test statistic and estimated λ by the median-based estimator. This test statistic is referred to as *GC-MED* in the following chapters. The analog test proposed by Reich and Goldstein (2001) based on the allelic χ^2-test and the mean-based estimator for λ is denoted by *GC-MEAN*. To calculate a conservative p-value Reich and Goldstein (2001) proposed to use the upper limit of the confidence interval for λ as a conservative estimate in the test statistic. Alternatively to their originally proposed method, Devlin et al. (2004) suggested to use the mean-based estimator in the test statistic but proposed the $F(1, L)$-distribution as an approximation

to the distribution of the test statistic. The F-distribution is the correct distribution for a quotient of χ^2-distributed variables and accounts for the variance when estimating λ, especially if the number of loci L is small. However, for $L \to \infty$ the limit distribution is equal to the χ^2_1-distribution and for a large number of loci the difference between both distributions is small.

4.2.2 The observed type-I error rate

In this section we would like to draw some conclusions from the distribution of the mean-based estimator $\widehat{\lambda}$. We investigate the influence of the distribution on the observed type-I error rate of Genomic Control for the *GC-MEAN* statistic.

Proposition 4.4. *An interval where the estimator $\widehat{\lambda}$ lies with a probability of $(1-\delta)$ given the true variance inflation λ is given by*

$$\left[\frac{\lambda}{L}\chi^2_{L;\delta/2}, \frac{\lambda}{L}\chi^2_{L;1-\delta/2}\right].$$

The $(1-\delta)$-confidence interval for the true variance inflation λ given the estimator $\widehat{\lambda}$ can be written as

$$\left[\frac{L\widehat{\lambda}}{\chi^2_{L;1-\delta/2}}, \frac{L\widehat{\lambda}}{\chi^2_{L;\delta/2}}\right].$$

Proof: The intervals can be directly calculated under the assumption that $\widehat{\lambda}$ is approximately $\frac{\lambda}{L}\chi^2_L$-distributed. □

A further step is to calculate an interval for the observed type-I error rate of GC which is observed if the estimator $\widehat{\lambda}$ is applied instead of the true variance inflation λ. We assume for these calculations that always the original estimator for λ is used, instead of changing the estimator to 1 for values smaller than 1. Let F_λ, F and $F_{\widehat{\lambda}}$ denote the distribution function for the $\lambda\chi^2_1$-,χ^2_1- and $\widehat{\lambda}\chi^2_1$-distribution respectively.

Proposition 4.5. *If the true variance inflation is equal to λ the expected type-I error rate of the allelic χ^2-test to the level α is given by*

$$1 - F\left(\frac{1}{\lambda}\chi^2_{1;1-\alpha}\right).$$

An interval which contains the observed type-I error rate of GC to the level α with a probability of $1-\delta$ can be calculated as

$$\left[1 - F\left(\frac{1}{L}\chi^2_{L;1-\delta/2}\chi^2_{1;1-\alpha}\right), 1 - F\left(\frac{1}{L}\chi^2_{L;\delta/2}\chi^2_{1;1-\alpha}\right)\right].$$

Proof: The relationship between the scaled χ^2-distribution and the simple χ^2-distribution is given by

$$F_\lambda(\lambda x) = F(x), \qquad F_\lambda^{-1}(\alpha) = \lambda F^{-1}(\alpha).$$

If the estimator $\widehat{\lambda}$ is applied instead of the true variance inflation λ the observed type-I error rate for GC to the level α can be calculated as

$$P(T^2 \geq F_{\widehat{\lambda}}^{-1}(1-\alpha)) = 1 - F_\lambda(F_{\widehat{\lambda}}^{-1}(1-\alpha)) = 1 - F\left(\frac{\widehat{\lambda}}{\lambda}\chi^2_{1;1-\alpha}\right).$$

For $\widehat{\lambda} = 1$ this formula gives the observed type-I error rate of the allelic χ^2-test. With the interval for $\widehat{\lambda}$ given in proposition 4.4 the GC interval can be calculated. □

This interval is only dependent on the number of loci used to estimate the variance inflation factor and not on the variance inflation factor itself because λ is cancelled out. Thus, the variation of the observed type-I error rate is neither influenced by the sample size of the study nor the amount of population structure. However, if GC is applied with a lower bound of $\widehat{\lambda} = 1$, the upper border of the interval for the observed type-I error rate of GC is bounded by the the expected type-I error rate for the χ^2-test which assumes no variance inflation. In this case GC always has a lower type-I error rate than the χ^2-test.

5 Structured Association

The second approach proposed in the literature to test for association in the presence of unknown population stratification is the method of Structured Association (SA). The model based Structured Association approach makes different use of the additionally genotyped marker loci. Population structure is directly inferred and the test of association incorporates the estimated population structure.
The concept of Structured Association originally has been developed by Pritchard et al. (2000a,b) and later on several different Structured Association methods were proposed (Satten et al., 2001; Zhu et al., 2002; Chen et al., 2003; Hoggart et al., 2003; Purcell and Sham, 2004). In this chapter we would like to introduce our own method of Structured Association (Köhler and Bickeböller, 2006) and describe the differences to existing methods. Here we give the new approach in detail with an emphasis on the comparison to other Structured Association approaches.
The Structured Association approaches can be mainly divided into two categories as one- and two-step approaches. Pritchard et al. (2000a,b); Purcell and Sham (2004) consider a two-step approach where the first step consists of modelling population structure for the entire study sample and the second step is the test of association based on the inferred structure. In contrast, Satten et al. (2001); Zhu et al. (2002); Chen et al. (2003); Hoggart et al. (2003) propose one-step approaches which simultaneously estimate population structure and test for association. Our procedure is also a two-step approach and belongs to the first category. Thus, based on the classification in one- and two-step approaches this chapter is structured as follows. In section 5.1 methods for clustering the individuals into subpopulations as the first step of Structured Association are discussed. Our clustering method is introduced and compared to other methods. In section 5.2 methods to test for association based on the inferred structure as the second step of Structured Association are described and our own test statistic is derived. In section 5.3 the one-step approaches for Structured Association are briefly described. Finally in section 5.4 the different methods of Structured Association are theoretically discussed.

5.1 Inference on population structure

Most of the Structured Association methods proposed so far have in common that they use a probability based approach for clustering the individuals into subpopulations. There are basically two concepts to apply a probability based approach. The more simple approach assumes that the population consists of discrete subpopulations. Then a mixture model can be applied and the subpopulations can be inferred via an EM algorithm. An EM algorithm was first applied by Purcell and Sham (2004) in a two-step approach. The second possibility to cluster the individuals is a Bayesian approach proposed by Pritchard et al. (2000a). The advantage of this approach is that it is possible to include admixed

individuals with alleles from more than one subpopulation. In this model a Markov Chain Monte Carlo (MCMC) algorithm can be used to infer population structure.
However, the two-step approaches proposed so far all have the disadvantage that they do not include phenotype information for clustering the individuals. They do not take into account that cases and controls a priori are expected to have a different probability of being from each of the subpopulations if confounding by population structure is suspected. We show here that even in a two-step approach it is necessary to include the information about the phenotype if the test statistic is based on the likelihood for the genotype data at the candidate locus. Otherwise the estimated subpopulation proportions in the case and the control group are biased leading to an inflated type-I error rate of the test statistic.
To determine the underlying genetic structure of a case-control sample we summarize the multilocus genotype marker data for L loci and N individuals in a vector $\boldsymbol{X}$ as introduced in section 2.4.

5.1.1 The standard mixture model

In this section we want to introduce the standard mixture model which we propose for clustering a population into discrete subpopulations if only multilocus genetic marker data and no other phenotypic information are available. For case-control data this standard mixture model has to be extended as described in section 5.1.2. Independently of our work, Purcell and Sham (2004) also proposed this model for clustering multilocus genetic marker data.
The standard mixture model is based on a discrete subpopulation model with fixed subpopulation allele frequencies (see section 2.3.2). Thus, we assume that the population consists of $k = 1, \ldots, K$ subpopulations. Within the subpopulations Hardy-Weinberg equilibrium (HWE) is assumed for each marker locus and linkage equilibrium between all marker loci. As described in section 2.3.2 and 2.4, the subpopulation allele frequencies are summarized in a vector $\boldsymbol{\varphi}$, the subpopulation proportions in the vector $\boldsymbol{\pi}$ and the unknown origin of each individual in a vector $\boldsymbol{Z}$. In the model here, $\boldsymbol{\varphi}$, $\boldsymbol{\pi}$ and $\boldsymbol{Z}$ and are unknown and have to be estimated.
Under the assumption of HWE, the genotype X_{il} of an individual from subpopulation k has a binomial distribution $B(2, \varphi_{kl})$ with probability mass function

$$f(x_{il}|k) = \binom{2}{x_{il}} \varphi_{kl}^{x_{il}} (1 - \varphi_{kl})^{2-x_{il}}$$

as described in section 2.2.1. The assumption of linkage equilibrium between the marker loci leads to a multivariate binomial distribution $B(2, \boldsymbol{\varphi}_k)$ for the complete genotype vector of an individual i with

$$f(\boldsymbol{x}_i|k) = \prod_{l=1}^{L} f(x_{il}|k).$$

Thus, the overall unconditional distribution of the genotype vector of an individual is a mixture of K multivariate binomial distributions. The mixture can then be described by the following equation

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_i|k).$$

Maximum likelihood estimators for the parameters of a mixture model $\boldsymbol{\varphi}$ and $\boldsymbol{\pi}$ can be determined for a given number of subpopulations applying an EM algorithm. The general form of the EM algorithm is described in the appendix A.2.2. The EM algorithm can be applied considering the genotype data as incomplete data and assuming the existence of the unobserved parameter vector $\boldsymbol{Z}$. The EM algorithm is chosen because the incomplete-data log-likelihood

$$\begin{aligned} \log L(\boldsymbol{\varphi}, \boldsymbol{\pi}|\boldsymbol{x}) &= \log P(\boldsymbol{x}) = \sum_{i=1}^{N} \log f(\boldsymbol{x}_i) \\ &= \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_i|k) \end{aligned}$$

is difficult to maximize since it contains the logarithm of a sum. In contrast, the complete-data log likelihood

$$\begin{aligned} \log L_C(\boldsymbol{\varphi}, \boldsymbol{\pi}|\boldsymbol{x}, \boldsymbol{z}) &= \log P(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{N} \log \left(f(\boldsymbol{x}_i|z_i) P(Z_i = z_i) \right) \\ &= \sum_{i=1}^{N} \log \left(\pi_{z_i} f(\boldsymbol{x}_i|z_i) \right) \end{aligned}$$

is easy to be maximized. However, the vector $\boldsymbol{Z}$ is unknown and hence an EM algorithm has to be applied. The following proposition gives the formulas for the EM algorithm in the mixture model described here.

Proposition 5.1 (standard EM algorithm)**.** *The EM algorithm for a mixture of K multivariate binomial distributions consists of the following two steps which have to be iteratively repeated for $t = 1, 2 \ldots$:*
E-step: The first step is to calculate the distribution of the unobserved data using the current parameter estimates $\boldsymbol{\varphi}^{(t)}, \boldsymbol{\pi}^{(t)}$ from iteration t, i.e.

$$q_{ik}^{*(t+1)} = P(Z_i = k|\boldsymbol{x}_i, \boldsymbol{\varphi}^{(t)}, \boldsymbol{\pi}^{(t)}) = \frac{\pi_k^{(t)} f(\boldsymbol{x}_i|k, \boldsymbol{\varphi}_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} f(\boldsymbol{x}_i|k', \boldsymbol{\varphi}_{k'}^{(t)})}.$$

M-step: In the second step the parameters are reestimated as

$$\begin{aligned}\pi_k^{(t+1)} &= \frac{1}{N}\sum_{i=1}^{N} q_{ik}^{*(t+1)},\\ \varphi_{kl}^{(t+1)} &= \frac{1}{2\sum_{i=1}^{N} q_{ik}^{*(t+1)}}\sum_{i=1}^{N} q_{ik}^{*(t+1)}\, x_{il}\,.\end{aligned}$$

Proof: The distribution of the unobserved data is calculated applying Bayes' formula. In the second step of the EM algorithm the function $Q(\boldsymbol{\varphi},\boldsymbol{\pi}|\boldsymbol{\varphi}^{(t)},\boldsymbol{\pi}^{(t)})$ defined in appendix A.2.2 has to be maximized. Independent of the concrete distribution of the data it can be shown for mixture models in general (Bilmes, 1998) that the function Q simplifies to

$$Q(\boldsymbol{\varphi},\boldsymbol{\pi}|\boldsymbol{\varphi}^{(t)},\boldsymbol{\pi}^{(t)}) = \sum_{k=1}^{K}\sum_{i=1}^{N} q_{ik}^{*(t+1)}\log\pi_k + \sum_{k=1}^{K}\sum_{i=1}^{N} q_{ik}^{*(t+1)}\log f(\boldsymbol{x}_i|k,\boldsymbol{\varphi}_k).$$

The two terms of the sum can be maximized separately. The first term has to be maximized with respect to $\boldsymbol{\pi}$ under the constraint that $\sum_{k=1}^{K}\pi_k = 1$ and this leads to the given formula for $\pi_k^{(t+1)}$ (Bilmes, 1998). The second term depends on the concrete distribution and has to be maximized with respect to $\boldsymbol{\varphi}$. For the multivariate binomial distribution it takes the form

$$\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{N} q_{ik}^{*(t+1)}\left[\log\binom{2}{x_{il}} + x_{il}\log\varphi_{kl} + (2-x_{il})\log(1-\varphi_{kl})\right].$$

The formula for $\varphi_{kl}^{(t+1)}$ follows by solving the equation

$$\frac{\partial}{\partial\varphi_{kl}} Q(\boldsymbol{\varphi},\boldsymbol{\pi}|\boldsymbol{\varphi}^{(t)},\boldsymbol{\pi}^{(t)}) = \sum_{i=1}^{N} q_{ik}^{*(t+1)}\left[\frac{x_{il}}{\varphi_{kl}} - \frac{2-x_{il}}{1-\varphi_{kl}}\right] = 0.$$

□

5.1.2 The mixture model for case-control data

For case-control data the additional information about the phenotype is available. We would like to show here that it is necessary to include this information in the clustering step to estimate the subpopulation proportions within case- and control group without a systematic bias. In a correct model it has to be assumed a priori that the distribution of the subpopulations is different between case and control group, i.e. in the mixture model the mixture proportions for cases and controls have to be modelled separately. The standard mixture model is extended by modelling two different mixture proportions $\boldsymbol{\pi}^{(y)} = (\pi_1^{(y)},\ldots,\pi_K^{(y)})'$ for $y = a$ (cases) and $y = u$ (controls) as in section 3.3. The vector

$\boldsymbol{Y}$ again contains the complete phenotype information of all individuals. The mixture can then be described by the equation

$$f(\boldsymbol{x}_i|y_i) = \sum_{k=1}^{K} P(Z_i = k|Y_i = y_i) f(\boldsymbol{x}_i|k, y_i) = \sum_{k=1}^{K} \pi_k^{(y_i)} f(\boldsymbol{x}_i|k)$$

since the distribution of the genotype data conditional on the subpopulation is independent of the phenotype. The EM algorithm has to be modified considering the genotype data conditional on the phenotype data as incomplete data and assuming again the existence of the unobserved parameter vector $\boldsymbol{Z}$. The incomplete-data log-likelihood can be written as

$$\log L(\boldsymbol{\varphi}, \boldsymbol{\pi}^{(a)}, \boldsymbol{\pi}^{(u)}|\boldsymbol{x}, \boldsymbol{y}) \quad = \quad \log P(\boldsymbol{x}|\boldsymbol{y}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k^{(y_i)} f(\boldsymbol{x}_i|k)$$

and the complete-data log likelihood has the form

$$\begin{aligned} \log L_C(\boldsymbol{\varphi}, \boldsymbol{\pi}^{(a)}, \boldsymbol{\pi}^{(u)}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &= \log P(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) = \sum_{i=1}^{N} \log \left(f(\boldsymbol{x}_i|z_i) P(Z_i = z_i|y_i) \right) \\ &= \sum_{i=1}^{N} \log \left(\pi_{z_i}^{(y_i)} f(\boldsymbol{x}_i|z_i) \right). \end{aligned}$$

The following proposition gives the formulas for the EM algorithm in the mixture model for case-control data. This EM algorithm is referred to as phenotype-dependent EM (P-EM).

Proposition 5.2 (phenotype-dependent EM algorithm)**.** *The EM algorithm for a mixture of K multivariate binomial distributions with different mixing proportions for cases and controls consists of the following two steps which have to be iteratively repeated for $t = 1, 2 \ldots$:*
E-step: The first step is to calculate the distribution of the unobserved data using the current parameter estimates $\boldsymbol{\varphi}^{(t)}, \boldsymbol{\pi}^{(a)(t)}, \boldsymbol{\pi}^{(u)(t)}$ from iteration t, i.e.

$$q_{ik}^{(t+1)} = P(Z_i = k|\boldsymbol{x}_i, y_i, \boldsymbol{\varphi}^{(t)}, \boldsymbol{\pi}^{(a)(t)}, \boldsymbol{\pi}^{(u)(t)}) = \frac{\pi_k^{(y_i)(t)} f(\boldsymbol{x}_i|k, \boldsymbol{\varphi}_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(y_i)(t)} f(\boldsymbol{x}_i|k', \boldsymbol{\varphi}_{k'}^{(t)})}.$$

M-step: In the second step the parameters are reestimated as

$$\begin{aligned} \pi_k^{(y)(t+1)} &= \frac{1}{N^{(y)}} \sum_{i:Y_i=y} q_{ik}^{(t+1)} \quad \textit{for} \quad y = a, u \\ \varphi_{kl}^{(t+1)} &= \frac{1}{2\sum_{i=1}^{N} q_{ik}^{(t+1)}} \sum_{i=1}^{N} q_{ik}^{(t+1)} x_{il}. \end{aligned}$$

Proof: The distribution of the unobserved data is calculated applying Bayes' formula conditional on the phenotype in the same way as for calculating the mixture distribution.

Adapting the proof of Bilmes (1998) we show that in the extended model the function Q (see appendix A.2.2) can be simplified similarly as before to

$$
\begin{aligned}
&Q(\boldsymbol{\varphi}, \boldsymbol{\pi}^{(a)}, \boldsymbol{\pi}^{(u)} | \boldsymbol{\varphi}^{(t)}, \boldsymbol{\pi}^{(a)(t)}, \boldsymbol{\pi}^{(u)(t)}) \\
&\quad = \sum_{\boldsymbol{z}} \log L_C(\boldsymbol{\varphi}, \boldsymbol{\pi}^{(a)}, \boldsymbol{\pi}^{(u)} | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) P(\boldsymbol{z} | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\varphi}^{(t)}, \boldsymbol{\pi}^{(a)(t)}, \boldsymbol{\pi}^{(u)(t)}) \\
&\quad = \sum_{(z_1,\ldots,z_N)} \sum_{i=1}^{N} \log \left(\pi_{z_i}^{(y_i)} f(\boldsymbol{x}_i | z_i, \boldsymbol{\varphi}_{z_i}) \right) \prod_{i'=1}^{N} q_{i' z_{i'}}^{(t+1)} \\
&\quad = \sum_{i=1}^{N} \sum_{z_i=1}^{K} \log \left(\pi_{z_i}^{(y_i)} f(\boldsymbol{x}_i | z_i, \boldsymbol{\varphi}_{z_i}) \right) q_{i z_i}^{(t+1)} \sum_{(z_1,\ldots,z_{i-1},z_{i+1},\ldots z_N)} \prod_{i':i' \neq i} q_{i' z_{i'}}^{(t+1)} \\
&\quad = \sum_{k=1}^{K} \sum_{i=1}^{N} \log \left(\pi_{k}^{(y_i)} f(\boldsymbol{x}_i | k, \boldsymbol{\varphi}_{k}) \right) q_{ik}^{(t+1)} \prod_{i':i' \neq i} \underbrace{\sum_{z_{i'}=1}^{K} q_{i' z_{i'}}^{(t+1)}}_{=1} \\
&\quad = \sum_{k=1}^{K} \sum_{i=1}^{N} q_{ik}^{(t+1)} \log \pi_{k}^{(y_i)} + \sum_{k=1}^{K} \sum_{i=1}^{N} q_{ik}^{(t+1)} \log f(\boldsymbol{x}_i | k, \boldsymbol{\varphi}_k) \\
&\quad = \sum_{k=1}^{K} \sum_{i:Y_i=a}^{N} q_{ik}^{(t+1)} \log \pi_{k}^{(a)} + \sum_{k=1}^{K} \sum_{i:Y_i=u}^{N} q_{ik}^{(t+1)} \log \pi_{k}^{(u)} + \sum_{k=1}^{K} \sum_{i=1}^{N} q_{ik}^{(t+1)} \log f(\boldsymbol{x}_i | k, \boldsymbol{\varphi}_k).
\end{aligned}
$$

Thus, the sum consists of three terms which can be maximized separately. The first two terms have to be maximized with respect to $\boldsymbol{\pi}^{(y)}$ under the constraint that $\sum_{k=1}^{K} \pi_k^{(y)} = 1$ for $y = a, u$ and this leads to the given formula for $\pi_k^{(y)(t+1)}$ in the same way as before (Bilmes, 1998). The third term which has to be maximized with respect to $\boldsymbol{\varphi}$ is the same as before and hence dependent on the genotype data of all individuals. Thus, for $\boldsymbol{\varphi}^{(t+1)}$ the same formula as before is obtained. □

5.1.3 The bias of the standard EM algorithm

The two-step approaches proposed so far for Structured Association (Pritchard et al., 2000a,b; Purcell and Sham, 2004) do not take phenotype information into account for clustering the individuals. We want to show that applying the standard EM algorithm for a case-control sample leads to asymptotically biased estimators of the subpopulation proportions within cases and controls. The final estimates of the standard EM algorithm are denoted as $q_{ik}^* = P(Z_i = k | \boldsymbol{x}_i)$ in comparison to the estimates of the P-EM algorithm $q_{ik} = P(Z_i = k | \boldsymbol{x}_i, y)$ for an individual with phenotype y. Applying Bayes' formula the relationship between those two parameters is given by

$$
q_{ik} = \frac{\gamma_k^{(y)} q_{ik}^*}{\sum_{k'=1}^{K} \gamma_{k'}^{(y)} q_{ik'}^*}
$$

where $\gamma_k^{(y)} = P(Y_i = y|Z_i = k)$ is the risk for phenotype y in subpopulation k within the study sample which does not depend on the marker data $\boldsymbol{x}_i$. We now specially want to consider the case $K = 2$ and characterize the bias introduced when estimating subpopulation proportions within cases and controls from the result of the standard EM algorithm, i.e. the bias in the estimator

$$\widehat{\pi}_k^{*(y)} = \frac{1}{N^{(y)}} \sum_{i:Y_i=y} q_{ik}^* \,.$$

Proposition 5.3. *For $K = 2$ the absolute difference between the subpopulation proportions in the case and control group is asymptotically systematically underestimated with the standard EM algorithm for $N \to \infty$, but $L < \infty$, i.e.*

$$\mathrm{E}\,|\widehat{\pi}_k^{*(a)} - \widehat{\pi}_k^{*(u)}| < |\pi_k^{(a)} - \pi_k^{(u)}| \quad \textit{for} \quad N \to \infty, L < \infty.$$

This holds as long as $\pi_k^{(a)} \neq \pi_k^{(u)}$.

Proof: The relationship between q_{i1} and q_{i1}^* simplifies for $K = 2$ to

$$q_{i1} = \frac{q_{i1}^*}{q_{i1}^* + \tau^{(y)}(1 - q_{i1}^*)}$$

where $\tau^{(y)} = \gamma_2^{(y)}/\gamma_1^{(y)}$. The three cases $0 < \tau^{(y)} < 1$, $\tau^{(y)} = 1$, $\tau^{(y)} > 1$ have to be distinguished. First, we want to concentrate on the case $\tau^{(y)} > 1$. The risk $\gamma_k^{(y)}$ can be written as a function of the true subpopulation proportions for phenotype y in comparison to the overall sample

$$\gamma_k^{(y)} = \frac{\pi_k^{(y)} P(Y = y)}{\pi_k}$$

and hence

$$\tau^{(y)} = \frac{\pi_2^{(y)}\pi_1}{\pi_2\pi_1^{(y)}} = \frac{(1 - \pi_1^{(y)})\pi_1}{(1 - \pi_1)\pi_1^{(y)}}.$$

Thus, $\tau^{(y)} > 1$ is equivalent to $\pi_1^{(y)} < \pi_1$ and hence $\pi_2^{(y)} > \pi_2$. For $\tau^{(y)} > 1$ the denominator of the above formula for q_{i1} is ≥ 1 which means that $q_{i1} \leq q_{i1}^*$ for all individuals i of phenotype y. Equality is only reached if $q_{i1}^* = 1$ or $q_{i1}^* = 0$. Hence, it follows $\widehat{\pi}_1^{(y)} < \widehat{\pi}_1^{*(y)}$ because there is some uncertainty in the estimates q_{i1}^* for $L < \infty$ and they are not all equal to 0 and 1. Since $\widehat{\pi}_1^{(y)}$ is the maximum likelihood estimator from the mixture model for case-control data the estimator is asymptotically unbiased for $N \to \infty$. Thus, $\widehat{\pi}_1^{*(y)}$ is asymptotically biased because the difference $\widehat{\pi}_1^{*(y)} - \widehat{\pi}_1^{(y)}$ does not converge to 0 for $N \to \infty$, $L < \infty$. If $\tau^{(y)} > 1$ and hence $\pi_1^{(y)} < \pi_1$ the unconditional estimate $\widehat{\pi}_1^{*(y)}$ systematically overestimates the true proportion $\pi_1^{(y)}$, i.e. $\pi_1^{(y)} < \mathrm{E}\,\widehat{\pi}_1^{*(y)} < \pi_1$. Analogously for $0 < \tau^{(y)} < 1$ the estimator $\widehat{\pi}_1^{*(y)}$ systematically underestimates the true proportion $\pi_1^{(y)}$, i.e. $\pi_1^{(y)} > \mathrm{E}\,\widehat{\pi}_1^{*(y)} > \pi_1$. The same holds for subpopulation 2 because subpopulation proportions sum up to 1. In conclusion, the estimators $\widehat{\pi}_1^{*(y)}$ and $\widehat{\pi}_2^{*(y)}$

are asymptotically biased towards the overall subpopulation proportions π_1 and π_2 and do not differ so much from them as the true subpopulation proportions $\pi_1^{(y)}$ and $\pi_2^{(y)}$ for phenotype y. Thus, the difference of the subpopulation proportions in case and control group is underestimated. Only for $\tau = 1$, i.e. $\pi_1^{(y)} = \pi_1$ and $\pi_2^{(y)} = \pi_2$ no bias appears. □

The underestimation of the differences in the subpopulation proportions between case and control group is expected to lead to an increased type-I error rate in the subsequent association test. In general, the bias becomes larger with an increasing difference of disease risks in the subpopulations as well as an increasing uncertainty in the classification into the subpopulations.

5.1.4 Extensions of the EM algorithm

In this section two extensions of the EM algorithm are discussed, first how to incorporate missing values and second how to handle multiple alleles. These extensions can be applied for both algorithms, the standard EM algorithm and the phenotype-dependent EM algorithm.

Missing values: Both EM algorithms are described for the perfect situation that the genotype data are fully available for all individuals and marker loci. However, in reality, the genotyping success is never equal to 100%. The algorithm has to be slightly changed to incorporate missing values. Let ν_{il} be an indicator variable which indicates if x_{il} is available or missing,

$$\nu_{il} = \begin{cases} 1 & : \quad x_{il} \text{ is available} \\ 0 & : \quad x_{il} \text{ is missing} \end{cases} .$$

In the E-step the distribution of the genotype data vector for each individual i is used to compute $q_{ik}^{*(t+1)}$ for the standard EM algorithm or $q_{ik}^{(t+1)}$ for the phenotype-dependent EM algorithm. In the case of missing values the random vector of the genotype data $\boldsymbol{X}_i$ only refers to all marker loci which are genotyped for individual i. The probability function for $\boldsymbol{X}_i$ (see section 5.1.1) can then be written as

$$f(\boldsymbol{x}_i|k) = \prod_{l:\nu_{il}=1} f(x_{il}|k).$$

The second change is in the allele frequency estimate calculated in the M-step of the EM algorithm. In the case of missing genotypes the estimate $\varphi_{kl}^{(t+1)}$ only refers to all individuals which are genotyped at locus l, i.e.

$$\varphi_{kl}^{(t+1)} = \frac{1}{2\sum_{i:\nu_{il}=1} q_{ik}^{(t+1)}} \sum_{i:\nu_{il}=1} q_{ik}^{(t+1)}\, x_{il} \,.$$

Multiple alleles: The other extension which shall be briefly mentioned is for multiallelic marker data especially microsatellites as described in section 2.2.4. Here again the distribution of the genotype data changes which affects the E-step of the EM algorithm. Under the assumption of HWE in the subpopulations the genotype data vector $\boldsymbol{X}_{il} = (X_{il}^{(1)}, \ldots, X_{il}^{(R_l)})'$ is multinomial$(2, \boldsymbol{\varphi}_{kl})$-distributed with subpopulation allele frequencies $\boldsymbol{\varphi}_{kl} = (\varphi_{kl}^{(1)}, \ldots, \varphi_{kl}^{(R_l)})'$. The probability function of $\boldsymbol{X}_{il}$ is given by

$$f(\boldsymbol{x}_{il}|k) = \frac{2}{\prod_{r=1}^{R_l} x_{il}^{(r)}!} \prod_{r=1}^{R_l} (\varphi_{kl}^{(r)})^{x_{il}^{(r)}}.$$

The multilocus genotype vector $\boldsymbol{X}_i$ then has a multivariate multinomial distribution with probability function

$$f(\boldsymbol{x}_i|k) = \prod_{l=1}^{L} f(\boldsymbol{x}_{il}|k)$$

under the assumption of linkage equilibrium between the marker loci as before.
In the M-step of the EM algorithm the allele frequencies have to be estimated as

$$\varphi_{kl}^{(r)(t+1)} = \frac{1}{2\sum_{i=1}^{N} q_{ik}^{(t+1)}} \sum_{i=1}^{N} q_{ik}^{(t+1)}\, x_{il}^{(r)}$$

for $r = 1, \ldots, R_l$.

5.1.5 Estimation of the number of subpopulations

A remaining problem is how to estimate the number of subpopulations. First of all, the EM algorithm has to be applied for several numbers of possible subpopulations K and the incomplete-data likelihood has to be calculated for the maximum-likelihood estimates, either

$$L_{\max}(K) = L(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\pi}}|\boldsymbol{x}, K)$$

for the standard EM algorithm or

$$L_{\max}(K) = L(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\pi}}^{(a)}, \widehat{\boldsymbol{\pi}}^{(u)}|\boldsymbol{x}, \boldsymbol{y}, K)$$

for the phenotype-dependent EM algorithm. An obvious way would be to use the likelihood ratio test for testing the null hypothesis that the population consists of K subpopulations versus the alternative hypothesis that the number of subpopulations is equal to $K + 1$. However, it is well known (Titterington et al., 1985) that the test statistic does not have the usual χ^2-distribution under the null hypothesis since regularity conditions do not hold for mixture models. The reason is that the null hypothesis lies on the boundary of the alternative hypothesis because the null hypothesis can be written in the way that an additional "dummy" population $K + 1$ exists with the parameter π_{K+1} equal to zero.

Therefore a parametric bootstrap procedure is recommended to estimate the distribution of the likelihood ratio test statistic (Böhning, 2000). The unknown parameters have to be estimated from the original data under the null hypothesis. Then bootstrap samples are drawn from a population with the estimated parameters. For these bootstrap samples the unknown parameters are estimated under the null hypothesis and under the alternative to determine the empirical distribution of the likelihood ratio test statistic.
However, in praxis there are also other criteria used to make computations feasible, e.g. the *Akaike information criterion* (*AIC*) which is based on a penalized likelihood (McLachlan and Peel, 2000). It must be noted, though, that the AIC is theoretically not justified since it is derived under the same regularity conditions as the likelihood ratio test statistic (Titterington et al., 1985). The AIC is defined as

$$\mathrm{AIC}(K) = -2 \log L_{\max}(K) + 2n(K)$$

where $n(K)$ is the number of free parameters in the mixture model. In the diallelic case the allele frequency vector $\boldsymbol{\varphi}$ has $L \cdot K$ free parameters and the vectors of the mixture proportions $\boldsymbol{\pi}, \boldsymbol{\pi}^{(a)}, \boldsymbol{\pi}^{(u)}$ each have $K - 1$ free parameters. Thus, for the standard EM algorithm the number of parameters is $n(K) = L \cdot K + K - 1$ and for the phenotype-dependent EM algorithm $n(K) = L \cdot K + 2(K - 1)$. The number of subpopulations with minimal $\mathrm{AIC}(K)$ is chosen. In practice, however, some problems have to be handled. Depending on the starting values the EM algorithm converges to different local maxima leading to different values of the log-likelihood. Thus, we propose to run the EM algorithm with different starting values for the same K. It must be ensured that finally a run of the EM algorithm is taken which does not converge to some local maximum with a rather small log-likelihood in comparison to the global maximum. How we exactly choose K in our simulations is described later in the simulation chapter.

5.1.6 The EM algorithm with admixture of Purcell and Sham (2004)

In real populations admixture between subpopulations is observed (see section 2.3.4). Purcell and Sham (2004) introduced a discrete model of admixture where parameters also can be inferred via EM algorithm. Their model only accounts for mixture LD but not for admixture LD between linked loci (see section 2.3.4). Purcell and Sham (2004) proposed the EM algorithm incorporating admixture as an extension of the standard EM algorithm but their model also can be applied for the phenotype-dependent EM algorithm. Thus, we would like to describe it in more detail. Admixture is modelled in terms of a finite number D of *derived classes* that represent an admixture of one or more of the K *ancestral subpopulations*. Thus, individuals of the same admixture proportions $\boldsymbol{q}_i^A$ are summarized in a derived class d. A derived class is described by the admixture proportions $\boldsymbol{\zeta}_d = (\zeta_{d1}, \ldots, \zeta_{dK})$ of its individuals where $\zeta_{dk} = q_{ik}^A$ is the proportion of the genome

originated from the ancestral subpopulation k for an individual i from the derived class d. The sum of the entries $\zeta_{dk}, k = 1, \ldots K$ is equal to 1. Derived classes are considered in a $1/S$ resolution where S is specified by the investigator. All possible derived classes are considered where each admixture proportion is a multiple of $1/S$. Thus, the derived classes also include pure classes which are only derived from one subpopulation. The unobserved data vector consists of two parts, one for the derived classes and the other for the ancestral subpopulations. For each individual Z_i^{DC} denotes the derived class (DC) of individual i. Furthermore, for each locus l and strand j the ancestral subpopulation Z_{ilj}^A where allele X_{ilj} is from is unknown. The EM algorithm is based on the relationship between derived classes and ancestral subpopulations. The E-step of the EM algorithm refers to calculating the posterior derived class probabilities of each individual. In the M-step the derived class proportions are either calculated for the whole sample (standard EM) or for cases and controls separately (phenotype-dependent EM). However, the allele frequency estimates are calculated for ancestral subpopulations. Thus, for each allele the posterior ancestral subpopulation probabilities have to be calculated based on the posterior derived class probabilities.

5.1.7 The Bayesian admixture model of Pritchard et al. (2000a)

The most popular method for inferring population structure is a Bayesian approach introduced by Pritchard et al. (2000a) and further developed by Falush et al. (2003). The method is implemented in the program STRUCTURE which has been widely used during the last years (see section 6.1.1). Pritchard et al. (2000a) first proposed a Bayesian approach for a similar discrete subpopulation model as described before. Thus, the same parameters are estimated but the estimation method is different. The advantage of the Bayesian approach is that it is possible to extend the model and to incorporate admixture in a continuous form. It seems to be a rather realistic model of population structure if the parameters can be correctly estimated. However, it is again not optimal for case-control data since the phenotype is not taken into account. In the following paragraph this model is described in more detail indicating how phenotype information could be incorporated in the model.

In the admixture model each individual i is assumed to have inherited some unknown proportion q_{ik}^A of its ancestry from each population k. As in the model of Purcell and Sham (2004) Z_{ilj}^A denotes the ancestral subpopulation where allele $\boldsymbol{X}_{ilj}$ is from. The vector notation $\boldsymbol{X}_{ilj}$ is used because the model is formulated for the multiallelic case. In the subpopulations again Hardy-Weinberg equilibrium and linkage equilibrium are assumed. Hence, in the terminology of Falush et al. (2003) background LD between tightly linked markers is not allowed (see section 2.3.4). Thus, conditional on the subpopulation it holds that

$$P(\{X_{ilj}^{(r)} = 1, X_{ilj}^{(r')} = 0 \text{ for all } r' \neq r\} | Z_{ilj}^A = k) = \varphi_{kl}^{(r)}$$

independently for each strand j and individual i.
There are different possibilities to define the distribution for the unknown subpopulations Z^A_{ilj}. The originally proposed model of Pritchard et al. (2000a) only models mixture LD, admixture LD is not included. Under the assumption of no admixture LD the ancestral subpopulations Z^A_{ilj} are independent for each locus l and strand j and have the a-priori distribution

$$P(Z^A_{ilj} = k) = q^A_{ik}.$$

The extended model proposed by Falush et al. (2003) additionally includes admixture LD. The ancestral subpopulations Z^A_{ilj} are dependent along each chromosomal strand, forming a Markov chain. It is assumed that chunks of chromosomes are derived as intact units from one of the subpopulations. Breakpoints between successive chunks occur at random and are modelled via a Poisson process. The subpopulation of origin of each chunk in individual i is independently drawn according to the vector $\boldsymbol{q}^A_i$. Because the ancestral subpopulations are unknown, the model is a hidden Markov model for the observed genotype data.
Estimation is performed in a Bayesian framework. The joint posterior distribution of the unobserved parameters $\boldsymbol{Z}^A, \boldsymbol{\varphi}, \boldsymbol{q}^A$ given the genotype data $\boldsymbol{X}$ must be inferred. Here the introduced parameters are summarized in vectors referring to the total sample. An MCMC algorithm is implemented to draw a sample from the joint posterior distribution. General information on constructing MCMC algorithms can be found in Gilks et al. (1996), for example. Prior distributions have to be specified for the admixture proportions and allele frequencies. In the updated version of Falush et al. (2003) the admixture proportions $\boldsymbol{q}^A_i$ are a priori independently Dirichlet($\boldsymbol{\alpha}$)-distributed where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$. The Dirichlet distribution is a generalization of the beta distribution to the multivariate case. It is very convenient to use the Dirichlet distribution for modelling admixture because a property of the Dirichlet distribution is that the entries of the vector $\boldsymbol{q}^A_i$ sum up to one. The expectation of the Dirichlet distribution is $\mathrm{E}\, q^A_{ik} = \alpha_k/\alpha_0$ where $\alpha_0 = \sum_k \alpha_k$. If the expectation is denoted by $\pi_k = \mathrm{E}\, q^A_{ik}$, the coefficient π_k gives the probability of an allele being originated in subpopulation k and thus refers to alleles instead of individuals. Consequently, the Dirichlet distribution could be reparameterized as $\boldsymbol{\alpha} = (\alpha_0\pi_1, \ldots, \alpha_0\pi_K)'$ where the parameter α_0 determines the amount of admixture in the sample. Small values of the admixture parameter α_0 indicate only a small amount of admixture with the majority of individuals having most of their alleles from only one subpopulation. For a case-control sample this prior distribution could be changed. Cases could have a different prior distribution than controls, i.e. conditionally on the phenotype y the admixture proportions $\boldsymbol{q}^A_i$ could assumed to be a priori independently Dirichlet($\boldsymbol{\alpha}^{(y)}$)-distributed where $\boldsymbol{\alpha}^{(y)} = (\alpha^{(y)}_0\pi^{(y)}_1, \ldots, \alpha^{(y)}_0\pi^{(y)}_K)'$. The parameters of the distributions have to be estimated including a Metropolis-Hastings step in the MCMC algorithm.
For the allele frequencies a prior distribution has to be specified as well. Pritchard et al.

(2000a) proposed a Dirichlet$(1,\ldots,1)$ prior of dimension R_l to model the allele frequencies $\boldsymbol{\varphi}_{kl}$ at each locus l within each subpopulation k. The special form of the Dirichlet distribution with all parameters equal to one corresponds to a uniform distribution of the allele frequencies. This is a model where the allele frequencies in the subpopulations are independent which is suitable for large population structure. A correlated subpopulation allele frequency model that accounts for correlations between the allele frequencies in closely related subpopulations should be applied as prior model for smaller population structure. In this model the allele frequencies in a hypothetical ancestral population are assumed to have uniform priors and the prior distribution of the subpopulation allele frequencies is a conditional distribution based on the ancestral allele frequencies and the distance of the subpopulations to the ancestral population. The correlated subpopulation allele frequency model can be seen as an extension of the beta-binomial model (see section 2.3.2) which we will apply later for simulating diallelic multilocus marker data for discrete subpopulations.

5.2 Association tests based on the inferred structure

The second step of Structured Association is to test for association at the candidate locus based on the inferred structure. We want to focus on two association tests, the likelihood ratio test of Pritchard et al. (2000a) and the Wald test we propose. Both test statistics are based on the same likelihood function for the genotype data conditional on the phenotype data which is introduced first. Additionally, the concepts for an association test in a logistic regression model are summarized.

5.2.1 The likelihood function

The likelihood function for the genotype data at the candidate locus is based on the vector $\boldsymbol{q} = (\boldsymbol{q}_1', \ldots, \boldsymbol{q}_N')'$ which here in general describes the inferred structure conditional on the phenotype data. The vector $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{iK})'$ either contains the inferred posterior probabilities in the mixture model for case-control data (section 5.1.2) or the admixture proportions in the admixture model (section 5.1.7) for case-control data. It is also possible to use other algorithms than described before to infer population structure as long as their result is a vector $\boldsymbol{q}$ of the form described here. As before, the test statistics we consider are based on the likelihood for the genotype data at the candidate locus conditional on the phenotype data. To adjust for population structure the likelihood has to be additionally conditioned on the inferred structure. We want to emphasize again that the inferred structure itself has to be calculated conditional on the phenotype data as well since the likelihood is calculated conditional on the phenotype data.
For the diallelic candidate locus the notation introduced in chapter 3 is used, denoting the genotype data with the vector $\boldsymbol{G}$ and the allele frequencies at the candidate locus under the alternative with $\boldsymbol{p}^1 = (\boldsymbol{p}^{(a)'}, \boldsymbol{p}^{(u)'})'$ and simply with $\boldsymbol{p}$ under the null hypothesis.

Additionally, let $\boldsymbol{Z}^C = (Z_{11}^C, \ldots, Z_{N2}^C)'$ be the vector of the unknown subpopulations at the candidate locus where Z_{ij}^C denotes the subpopulation where the j-th allele of individual i at the candidate locus is from. Dependent on the vector $\boldsymbol{q}_i$ each individual i has its own allelic distribution. Under the alternative each allele G_{ij} is Bernoulli distributed with allele frequency

$$P_1(G_{ij} = 1|y_i, \boldsymbol{q}_i) = \sum_{k=1}^{K} P(Z_{ij}^C = k|y_i, \boldsymbol{q}_i) P(G_{ij} = 1|Z_{ij}^C = k, y_i) = \sum_{k=1}^{K} q_{ik} p_k^{(y_i)}.$$

It should be noted that for the second equality it is necessary that the admixture proportions are calculated conditionally on the phenotype. Under the null hypothesis the allele frequency in subpopulation k is independent of the disease status y_i. However, the distribution of allele G_{ij} is still dependent on the disease status y_i because the structure is calculated conditionally on the phenotype, i.e. the above formula simplifies under the null hypothesis to

$$P_0(G_{ij} = 1|y_i, \boldsymbol{q}_i) = \sum_{k=1}^{K} q_{ik} p_k.$$

The conditional distribution of G_i is a binomial distribution $B(2, \sum_{k=1}^{K} q_{ik} p_k^{(y_i)})$. Here an allele based test statistic is considered where the two alleles of one individual are allowed to be from different subpopulations and the subpopulations of the two alleles are considered as independent. This corresponds to the idea of the admixture model. We do not use the constraint that the two alleles of one individual should be from the same subpopulation if q_{ik} denotes the posterior probability in the mixture model. The likelihood function is then

$$\begin{aligned} L_1(\boldsymbol{p}^{(a)}, \boldsymbol{p}^{(u)}) &= P_1(\boldsymbol{g}|\boldsymbol{y}, \boldsymbol{q}) = \prod_{i:Y_i=a} f(g_i|a, \boldsymbol{q}_i) \prod_{i:Y_i=u} f(g_i|u, \boldsymbol{q}_i) \\ &= \prod_{i:Y_i=a} \binom{2}{g_i} \left(\sum_{k=1}^{K} q_{ik} p_k^{(a)} \right)^{g_i} \left(1 - \sum_{k=1}^{K} q_{ik} p_k^{(a)} \right)^{2-g_i} \\ &\quad \cdot \prod_{i:Y_i=u} \binom{2}{g_i} \left(\sum_{k=1}^{K} q_{ik} p_k^{(u)} \right)^{g_i} \left(1 - \sum_{k=1}^{K} q_{ik} p_k^{(u)} \right)^{2-g_i}. \end{aligned}$$

This simplifies under the null hypothesis to

$$L_0(\boldsymbol{p}) = P_0(\boldsymbol{g}|\boldsymbol{y}, \boldsymbol{q}) = \prod_{i=1}^{N} \binom{2}{g_i} \left(\sum_{k=1}^{K} q_{ik} p_k \right)^{g_i} \left(1 - \sum_{k=1}^{K} q_{ik} p_k \right)^{2-g_i}.$$

The log-likelihood $\log L_1(\boldsymbol{p}^1)$ has to be maximized with respect to $\boldsymbol{p}^{(a)}$ and $\boldsymbol{p}^{(u)}$ which leads to two separate systems of K non-linear equations, the first based on the genotype data for cases, the second based on the genotype data for controls. Maximizing the log-likelihood $\log L_0(\boldsymbol{p})$ for $\boldsymbol{p}$ leads to one system of K non-linear equations, based on the

complete genotype data. These systems again can be solved by an EM algorithm considering the genotype data $\boldsymbol{G}$ as incomplete and the vector $\boldsymbol{Z}^C$ as missing data. The following proposition gives the formula for the EM algorithm to maximize the log-likelihood under the alternative with respect to $\boldsymbol{p}^{(y)}$ and can be applied analogously for the log-likelihood under the null hypothesis.

Proposition 5.4. *The EM algorithm for maximizing the log-likelihood under the alternative with respect to* $\boldsymbol{p}^{(y)}$ *consists of the following steps which have to be iteratively repeated for* $t = 1, 2 \ldots$*:*
E-step: The first step is to calculate the distribution of the unobserved data using the current parameter estimates $\boldsymbol{p}^{(y)(t)}$ *from iteration* t*. It is sufficient to consider the individuals* i *at strand* j *where* $G_{ij} = 1$ *and* $Y_i = y$*.*

$$q_{ijk}^{C(t+1)} := P(Z_{ij}^C = k | G_{ij} = 1, Y_i = y, \boldsymbol{p}^{(y)(t)}) = \frac{q_{ik} p_k^{(y)(t)}}{\sum_{k'=1}^K q_{ik'} p_{k'}^{(y)(t)}}$$

M-step: In the second step the allele frequencies are reestimated as

$$p_k^{(y)(t+1)} = \frac{1}{\sum_{i:Y_i=y} \sum_{j=1,2} q_{ijk}^{C(t+1)}} \sum_{i:Y_i=y} \sum_{j=1,2} q_{ijk}^{C(t+1)} G_{ij} \,.$$

Proof: The EM algorithm can be derived in the same way as for the mixture model. □

The maximum likelihood estimates are denoted as $\widehat{\boldsymbol{p}}^{(a)}$ and $\widehat{\boldsymbol{p}}^{(u)}$. Two special cases require additional remarks. If discrete subpopulations are considered and subpopulation membership is known in advance, the admixture proportions are $q_{ik} = 1$ and $q_{ik'} = 0$ for $k' \neq k$ for an individual i from subpopulation k. The maximum-likelihood equations can be solved for each k independently leading to the usual maximum likelihood estimates. The second extreme case is the case where no subpopulation information is available for concrete individuals, only the subpopulation proportions are known within cases and controls. Then $q_{ik} = \pi_k^{(y)}$ for all individuals. In this case the maximum-likelihood equations cannot be uniquely solved.

5.2.2 The likelihood ratio test of Pritchard et al. (2000b)

Pritchard et al. (2000b) proposed a likelihood ratio test denoted as *STRAT* (Structured Association Test) which is based on the likelihoods calculated above. However, Pritchard et al. (2000b) proposed the test without conditioning the admixture proportions on the phenotype. This is theoretically not correct as discussed in the previous section. The test can be seen as an extension of the unrestricted likelihood ratio test described in proposition 3.8 for discrete subpopulations and known subpopulation membership. Pritchard et al.

(2000b) recommended to calculate the p-value by simulations. However, since the likelihood is calculated conditioning on the inferred structure the likelihood ratio test statistic has the usual asymptotic χ^2-distribution.

Proposition 5.5. *The unrestricted likelihood ratio test statistic based on the likelihoods calculated in section 5.2.1*

$$-2\log\left(\frac{L_0(\widehat{\boldsymbol{p}})}{L_1(\widehat{\boldsymbol{p}}^{(a)},\widehat{\boldsymbol{p}}^{(u)})}\right)$$

is asymptotically χ^2_K-distributed under the null hypothesis $H_0 : \boldsymbol{p}^{(a)} = \boldsymbol{p}^{(u)}$.

Proof: The asymptotic theory of likelihood ratio tests is described in the appendix A.2.1. Under the alternative K additional parameters have to be estimated. □

The likelihood ratio test has the advantage that the test statistic is easily calculated and that it can be generalized straight forward for the multiallelic case. However, as we discussed in section 3.3.2 the unrestricted likelihood ratio test is not optimal if population stratification only acts as a confounder and not as an effect modifier. Since a restricted likelihood ratio test has the disadvantage that the parameters have to be estimated under the restricted alternative we construct a Wald test for the situation that only confounding effects of population structure are expected and the allelic odds ratios are assumed to be the same in all subpopulations.

5.2.3 The Wald test

In this section a Wald test is derived which later is referred to as *CHIPOP*. To calculate a Wald-type test statistic, the variance of the maximum likelihood estimates has to be asymptotically calculated.

Proposition 5.6. *The asymptotic distribution for the maximum likelihood estimates is*

$$\sqrt{N}\left(\widehat{\boldsymbol{p}}^1 - \boldsymbol{p}^1\right) \stackrel{\cdot}{\sim} N\left(\boldsymbol{0}, N\boldsymbol{I}^{-1}(\boldsymbol{p}^1)\right)$$

The information matrix splits into two parts

$$\boldsymbol{I}(\boldsymbol{p}^1) = \operatorname{diag}\left(\boldsymbol{I}^{(a)}(\boldsymbol{p}^{(a)}), \boldsymbol{I}^{(u)}(\boldsymbol{p}^{(u)})\right)$$

where $\boldsymbol{I}^{(y)}(\boldsymbol{p}^{(y)})$ is a $K \times K$ matrix with the entries

$$[\boldsymbol{I}^{(y)}(\boldsymbol{p}^{(y)})]_{k,k*} = 2\sum_{i:Y_i=y} \frac{q_{ik}q_{ik^*}}{\sum_{k'=1}^K q_{ik'}p_{k'}^{(y)}\left(1-\sum_{k'=1}^K q_{ik'}p_{k'}^{(y)}\right)}.$$

Proof: The log-likelihood function $l_1 = \log L_1$ can be written as

$$l_1(\boldsymbol{p}^{(a)},\boldsymbol{p}^{(u)}) = \sum_{i:Y_i=a}\left[\log\binom{2}{g_i} + g_i \log\left(\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(a)}\right) + (2-g_i)\log\left(1-\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(a)}\right)\right] + \sum_{i:Y_i=u}\left[\log\binom{2}{g_i} + g_i \log\left(\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(u)}\right) + (2-g_i)\log\left(1-\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(u)}\right)\right].$$

For the first derivatives it follows that

$$\frac{\partial}{\partial p_k^{(y)}} l_1(\boldsymbol{p}^{(a)},\boldsymbol{p}^{(u)}) = \sum_{i:Y_i=y} g_i \frac{q_{ik}}{\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(y)}} - (2-g_i)\frac{q_{ik}}{1-\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(y)}}.$$

The second derivatives with respect to the same phenotypes are calculated as

$$\frac{\partial^2}{\partial p_k^{(y)}\partial p_{k^*}^{(y)}} l_1(\boldsymbol{p}^{(a)},\boldsymbol{p}^{(u)}) = \sum_{i:Y_i=y} -g_i \frac{q_{ik}q_{ik^*}}{\left(\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(y)}\right)^2} - (2-g_i)\frac{q_{ik}q_{ik^*}}{\left(1-\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(y)}\right)^2}.$$

All second derivatives with respect to discordant phenotypes are zero, i.e. for $y \neq y^*$

$$\frac{\partial^2}{\partial p_k^{(y)}\partial p_{k^*}^{(y^*)}} l_1(\boldsymbol{p}^{(a)},\boldsymbol{p}^{(u)}) = 0.$$

The entries of the part of the information matrix which corresponds to phenotype y can be calculated as

$$\begin{aligned}
[\boldsymbol{I}^{(y)}(\boldsymbol{p}^{(y)})]_{k,k*} &= -\mathrm{E}\left[\frac{\partial^2}{\partial p_k^{(y)}\partial p_{k^*}^{(y)}} l_1(\boldsymbol{p}^{(a)},\boldsymbol{p}^{(u)})\right] \\
&= \sum_{i:Y_i=y} 2\frac{q_{ik}q_{ik^*}}{\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(y)}} + 2\frac{q_{ik}q_{ik^*}}{1-\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(y)}} \\
&= 2\sum_{i:Y_i=y} \frac{q_{ik}q_{ik^*}}{\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(y)}\left(1-\sum_{k'=1}^{K} q_{ik'}p_{k'}^{(y)}\right)}.
\end{aligned}$$

The asymptotic normal distribution follows from the standard maximum likelihood theory as described in the appendix in proposition A.1. □

The Wald test we propose is based on the allele frequency differences between cases and controls averaged over the subpopulations. Let the vector $(c_1,\ldots,c_K)'$ contain the weights which are assigned to each allele frequency difference and let the contrast vector be $\boldsymbol{c} = (c_1,\ldots,c_K,-c_1,\ldots-c_K)'$.

Proposition 5.7. *Let $\widehat{\boldsymbol{\Sigma}}_N$ be the variance estimator under the null hypothesis $H_0 : \boldsymbol{p}^{(a)} = \boldsymbol{p}^{(u)}$, i.e*

$$\widehat{\boldsymbol{\Sigma}}_N = \mathrm{diag}\left(\boldsymbol{I}^{(a)}(\widehat{\boldsymbol{p}})^{-1}, \boldsymbol{I}^{(u)}(\widehat{\boldsymbol{p}})^{-1}\right)$$

based on the definitions of proposition 5.6. The Wald-type test statistic

$$(\boldsymbol{c}'\widehat{\boldsymbol{p}}^1)'[\boldsymbol{c}'\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{c}]^{-1}\boldsymbol{c}'\widehat{\boldsymbol{p}}^1 = \frac{\left[\sum_{k=1}^K c_k(\widehat{p}_k^{(a)} - \widehat{p}_k^{(u)})\right]^2}{\sum_{k=1}^K\sum_{k^*=1}^K c_k c_{k^*}\left([\boldsymbol{I}^{(a)}(\widehat{\boldsymbol{p}})^{-1}]_{kk^*} + [\boldsymbol{I}^{(u)}(\widehat{\boldsymbol{p}})^{-1}]_{kk^*}\right)}$$

is asymptotically χ^2_1-distributed for $N \to \infty$, but $\pi_k^{(a)} \geq s$ and $\pi_k^{(u)} \geq s$ for all $k = 1, \ldots, K$ and some constant $s > 0$, under the null hypothesis of no association in the subpopulations.

Proof: Under the null hypothesis of no association $H_0 : \boldsymbol{p}^{(a)} = \boldsymbol{p}^{(u)}$ it follows that

$$\boldsymbol{c}'\boldsymbol{p}^1 = \sum_{k=1}^K c_k(p_k^{(a)} - p_k^{(u)}) = 0$$

for each vector $\boldsymbol{c}$ as defined above. Thus, the Wald test statistic can be based on the weighted average $\boldsymbol{c}'\widehat{\boldsymbol{p}}^1$. The asymptotic variance $\boldsymbol{\Sigma}_N = \boldsymbol{I}(\boldsymbol{p}^1)^{-1}$ can be estimated under the null hypothesis of no association replacing $\boldsymbol{p}^{(a)}$ and $\boldsymbol{p}^{(u)}$ by $\widehat{\boldsymbol{p}}$. As described in proposition A.2 the Wald-type statistic is then asymptotically χ^2-distributed with 1 df because $\boldsymbol{c}$ is only a vector and hence of rank 1. For the asymptotic distribution it has to be assumed for each subpopulation k that $\pi_k^{(a)} \geq s$ and $\pi_k^{(u)} \geq s$. Thus, the convergence refers to the number of alleles from each subpopulation k within cases and controls, i.e. $N^{(a)}\pi_k^{(a)} \to \infty$ and $N^{(u)}\pi_k^{(u)} \to \infty$. □

The Wald test we propose here can be viewed as an extension of the Cochran-Mantel-Haenszel test described in proposition 3.6 which is the standard test adjusting for a discrete and known confounding variable. In the case of discrete subpopulations the test statistic simplifies to the Cochran-Mantel-Haenszel test.
In principle it would be possible to take any weights, for example the subpopulations could be equally weighted. However, we propose to take

$$c_k = \frac{2}{1/(N^{(a)}\widehat{\pi}_k^{(a)}) + 1/(N^{(u)}\widehat{\pi}_k^{(u)})}$$

which is the harmonic mean of the estimated number of alleles from subpopulation k within cases and controls. Thus, the variance of the test statistic is reduced because the weight of a subpopulation is low if the proportion of a subpopulation is small in cases or in controls. The weights are adapted from the Mantel-Haenszel test to the case of unknown subpopulations.
Based on the same arguments as in the case of discrete subpopulations it can be shown that the numerator of the test statistic

$$W = \sum_{k=1}^K c_k(\widehat{p}_k^{(a)} - \widehat{p}_k^{(u)})$$

is an asymptotically unbiased estimator for a common log odds ratio in all subpopulations k multiplied by a constant. Thus, the Wald test statistic should be applied in the case where a common odds ratio in the subpopulations is assumed. However, the weights do not fulfill any optimality criterion as in the discrete case where the weights are chosen to maximize the Pitman efficiency of the test statistic for the alternative hypothesis of a common odds ratio. Weights which are optimal in this sense could be constructed here as well but have a very complicated form dependent on the allele frequency estimates, thus we do not use them here. However, since the weights are adapted from the Mantel-Haenszel test they should be quite close to the optimal weights.

5.2.4 Logistic regression as an alternative approach

Alternatively, logistic regression could be applied to test for association conditionally on the inferred structure in the model without admixture. We decided to concentrate on the classical approaches to test for association as described before and just would like to outline the idea to build up a logistic regression model. The logistic regression is applied conditional on the subpopulation

$$\text{logit}\left(P(Y_i = a|g_i, Z_i = k)\right) = \alpha + \beta g_i + \eta_k$$

where β models the effect of the candidate gene and η_k the effect of subpopulation k as described in section 3.4. This is the same equation which Satten et al. (2001); Zhu et al. (2002) use in their one-step approach explained in section 5.3.1. The idea is to derive an expression for the likelihood $P(\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{q})$ of the phenotype data conditional on the genotype data at the candidate locus and the inferred structure. In this case, the vector of the inferred structure $\boldsymbol{q}$ contains the posterior probabilities of each individual belonging to each of the subpopulations estimated in a standard mixture model. These probabilities do not have to be calculated conditionally on the phenotype because the logistic model is based on the likelihood of the phenotype data. The likelihood for each individual i can be formulated as a mixture of K subpopulations

$$\begin{aligned} P(y_i|g_i, \boldsymbol{q}_i) &= \sum_{k=1}^{K} P(Z_i = k|g_i, \boldsymbol{q}_i) P(Y_i = y_i|g_i, Z_i = k) \\ &= \sum_{k=1}^{K} q_{ik} P(Y_i = y_i|g_i, Z_i = k) \end{aligned}$$

under the assumption that the probability that an individual belongs to a certain subpopulation is not dependent on the genotype data at the candidate locus. The total likelihood then is the product of the likelihoods for each individual. The probability $P(Y_i = y_i|g_i, Z_i = k)$ can be written as a function of the parameters α, η_k and β using the logistic regression model. Some technical work has to be carried out to derive an iterative

algorithm to determine the maximum likelihood estimates for the regression parameters and to construct the test statistics. Details could be worked out in the future.

5.3 One-step approaches

Satten et al. (2001); Zhu et al. (2002); Chen et al. (2003); Hoggart et al. (2003) proposed one-step approaches for Structured Association. The idea is to simultaneously estimate population structure and test for association at the candidate locus. For comparison with our approach the two most popular one-step approaches are briefly summarized. Although one-step approaches seem to be statistically superior they have disadvantages as discussed in detail in section 5.4.

5.3.1 The one-step mixture model of Satten et al. (2001)

Satten et al. (2001) proposed a one-step approach assuming that the total population is a mixture of K discrete subpopulations. Zhu et al. (2002) later suggested a similar method, thus the two approaches can be described together. To model structure and association simultaneously, the mixture is assumed to be also dependent on the genotype data at the candidate locus. Satten et al. (2001) consider a mixture of multivariate binomial distributions as described before. However, Zhu et al. (2002) do not consider the original marker data in the model but propose to start with a principal component analysis to reduce the dimensionality and consider a mixture of multivariate normal distributions. For the test of association a logistic regression model conditional on the subpopulation is proposed as described in section 5.2.4.
The idea of Satten et al. (2001) is to combine the mixture model and the logistic regression by deriving an expression for the combined likelihood $P(\boldsymbol{x}, \boldsymbol{g}|\boldsymbol{y})$ of the marker data and the genotype data at the candidate locus given the phenotype data. Zhu et al. (2002) instead consider the combined likelihood $P(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{g})$ of the marker data and the phenotype data given the genotype data at the candidate locus which is more straight forward if a logistic regression model is applied (see section 3.4). The likelihood for each individual i can be formulated as a mixture of K subpopulations

$$\begin{aligned} P(\boldsymbol{x}_i, y_i|g_i) &= \sum_{k=1}^{K} P(Z_i = k|g_i) P(\boldsymbol{x}_i, y_i|g_i, Z_i = k) \\ &= \sum_{k=1}^{K} P(Z_i = k) P(Y_i = y_i|g_i, Z_i = k) f(\boldsymbol{x}_i|Z_i = k), \end{aligned}$$

again under the assumption that the probability that an individual belongs to a certain subpopulation is not dependent on the genotype data at the candidate locus. The likelihood can be interpreted as a mixture likelihood where the mixture proportions

$P(Z_i = k)P(Y_i = y_i|g_i, Z_i = k)$ are dependent on the phenotype and genotype of the individual. The probability $P(Y_i = y_i|g_i, Z_i = k)$ can be written as a function of the parameters α, η_k and β using the logistic regression model. The maximum likelihood estimates for this model can be determined applying an EM algorithm. However, one disadvantage of the model is that hypothesis testing is not straight forward because the variance of the maximum likelihood estimate $\widehat{\beta}$ cannot be directly calculated due to the large number of parameters in the model. A likelihood ratio test could be applied additionally maximizing the likelihood under the null hypothesis $H_0 : \beta = 0$. However, Satten et al. (2001) recommend not to use this test statistic because different subpopulations could be inferred when repeating the algorithm under the null hypothesis. Satten et al. (2001) instead propose to apply a parametric bootstrap approach as described in section 5.1.5 to estimate the variance of the maximum likelihood estimate $\widehat{\beta}$.

5.3.2 The one-step Bayesian model of Hoggart et al. (2003)

Hoggart et al. (2003) developed a rather complicated but very general Bayesian approach implemented in the program ADMIXMAP. The program can be applied to analyze data sets which consist of a quantitative or binary trait and multilocus genotype data from a sample of individuals drawn from an admixed population. One main application is a case-control study where the disease status is the binary trait of interest. In contrast to the models explained before, the model does not make a difference between marker and candidate loci. Here, each marker locus used to infer population structure can be additionally tested for an association to the phenotype. Admixture is modelled in a Bayesian framework similar to the approach of Pritchard et al. (2000a); Falush et al. (2003) by specifying prior distributions for the model parameters and calculating the posterior distributions by an MCMC algorithm. The main difference to Pritchard et al. (2000a); Falush et al. (2003) is that admixture is modelled dependent on the phenotype but in a different way as we propose. A logistic regression model is fitted to model the dependence of the phenotype on individual admixture and possible covariates. The posterior distribution of the regression coefficients is part of the result of the MCMC algorithm. To test for association between a marker locus and the disease, the logistic regression model under the alternative is considered where the phenotype is additionally dependent on the genotype at the marker locus. A score test is derived to test if the regression parameter for the genotype at the marker locus is equal to zero. The score test has two advantages. The score (gradient of the log-likelihood) is calculated for each realization of the complete data and uncertainty about the admixture proportions is accounted for by including the posterior variance of the realized score in the score statistic. It is computationally efficient allowing all loci to be tested for association in a single run of the MCMC algorithm because parameters only have to be estimated under the null hypothesis of no association to calculate the score statistic. Thus, the model of Hoggart et al. (2003) has some advantages compared to other

models but the disadvantage of all Bayesian approaches as discussed subsequently.

5.4 Discussion

The approach of Pritchard et al. (2000a,b) is the most popular Structured Association approach. However, we showed that such an approach is theoretically not valid because Structured Association has to be applied with a clustering algorithm conditioning on the phenotype if subsequently a test statistic based on the likelihood function for the genotype data at the candidate locus is applied. Otherwise a systematic bias when estimating the subpopulation proportions within cases and controls is introduced which leads to an inflated type-I error rate of the test statistics. This point has not been identified as crucial before, although it has also been noted by others in their simulations that the Structured Association approach of Pritchard et al. (2000b) is too liberal if population structure cannot be inferred correctly (Zhu et al., 2002, table II). We claim that this can be explained by not taking phenotype information into account. In contrast, the approach of Satten et al. (2001); Zhu et al. (2002) is a one-step approach which correctly conditions on the phenotype. It has the disadvantage that for each candidate locus the sample has to be clustered again. First of all this is time consuming because for each candidate locus it has to be ensured that a proper run of the EM algorithm is finally chosen converging to the true maximum of the likelihood. Secondly, in the situation where several candidate loci and additionally some null loci have been genotyped an approach might be preferred where first the structure of the population is described based on the null loci and then several candidate genes are tested. As already mentioned the other disadvantage is that hypothesis testing is not straight forward because the variance of the effect estimate cannot be directly calculated due to the large number of parameters in the model. Thus, we showed that it is sufficient to apply a two-step approach with the clustering algorithm conditioning on the phenotype. Compared to Satten et al. (2001) we do not loose much information in a two-step approach because we use the posterior probabilities of each individual belonging to each of the subpopulations when testing for association.
For simplicity, we concentrate on the EM algorithm for clustering the individuals into discrete subpopulations. As already discussed, it is also possible to extend the idea of incorporating phenotype information in the clustering process to the Bayesian model of Pritchard et al. (2000a); Falush et al. (2003) where admixture of populations is allowed. Again, it should be noted that the matrix $\boldsymbol{q}$ which describes the inferred structure has to be interpreted carefully. Modelling discrete subpopulations the matrix $\boldsymbol{q}$ contains the posterior probabilities for each individual of being in each of the subpopulations. Uncertainty in the classification of the individuals due to a small number of null loci is therefore automatically accounted for because in the association test these posterior probabilities are used instead of assigning the individuals to the most likely subpopulation. However, if admixture is modelled in a Bayesian framework the matrix $\boldsymbol{q}$ describes the estimated

admixture proportions of each individual. Here, uncertainty in the classification is not included when applying the likelihood ratio test or the Wald test proposed in section 5.2. In contrast, the model of Hoggart et al. (2003) has the advantage that the uncertainty about the admixture proportions is accounted for when testing for association. However, the question remains how well admixture models work in practice. In principle, it is desirable to have a model where admixture of populations is allowed because in real populations admixture is always present. However, admixture models have a lot of parameters to be estimated and thus primarily Bayesian models are proposed to incorporate admixture. These Bayesian models require a lot of finetuning when specifying the prior distributions and the parameter estimates are very sensitive to the prior distributions. Moreover, it has to be assumed that even more loci have to be genotyped to model admixture close to reality. An additional disadvantage is that the MCMC algorithms are computationally very intensive which makes larger simulations impossible. As already shown a different possibility to model admixture could be to extend our phenotype-dependent EM algorithm to incorporate admixture in the form of derived classes as proposed by Purcell and Sham (2004).

A second question is which test statistic to use for Structured Association after applying the phenotype-dependent EM algorithm. Theoretically, the crucial point is if population stratification only acts as a confounder or additionally as an effect modifier. How large the differences between the two test statistics are in practice, is investigated later in simulations (see chapter 7).

6 Results from population based studies

In this chapter the impact of population stratification on case-control association studies is assessed in realistic situations of only subtle population stratification as within Europe or even within Germany. Results from recently published studies as well as results from the German Genomic Control Study are presented. Finally it is discussed if it seems to be feasible and reasonable to apply methods of Structured Association and Genomic Control for case-control studies within Europe or Germany.

6.1 Results from previous studies

6.1.1 The amount of stratification in real populations

Cavalli-Sforza et al. (1996) investigated the genetic history of world populations. Here we want to summarize some of their results to give an idea about the typical range of F_{ST}-values, especially within Europe. Based on original articles about genetic studies Cavalli-Sforza et al. (1996) collected gene frequency data from 491 world populations for around 120 loci. However, in the final analysis the number of populations had to be reduced to 42, pooling some populations and eliminating several less well tested populations. The largest genetic distance was detected between African and non-African populations with an average F_{ST}-value estimated at 0.205. The structure of the European continent turned out to be the most difficult to describe with a lot of genetic heterogeneity within European countries and small differences between them. In Europe finally 26 populations with 26.4% missing frequency data were compared at an average of 88 marker loci. The analysis shows that there are some outlier populations which are genetically distinct from the rest of Europe. The Lapps and Sardinians are the most extreme outliers, e.g. F_{ST}-values between Lapps and other European populations ranges from 0.021 to 0.067. Greeks, Yugoslavs, Basques, Icelanders and Finns are five less extreme outliers but the rest of Europe (central Europe) is fairly homogeneous. F_{ST}-values within central Europe range from close to 0 to 0.015. Within the subgroup of the Germanic populations comprising the Dutch, Danish, English, Austrians, Swiss, Germans, Belgians F_{ST}-values are estimated even smaller than 0.006. The F_{ST}-distance between Germans and one of these six populations lies between 0.0010 for Germans and Swiss and 0.0022 for Germans and English. However, these F_{ST}-values are only estimations based on a rather small number of genetic loci. Population samples from different genetic studies are mixed together not being uniquely genotyped. Thus, further studies are necessary to confirm these results.
In the last years the number of studies determining the genetic structure of human world populations increased with the improvement of the genotyping technology. Based on 60 to 400 marker loci the program STRUCTURE of Pritchard et al. (2000a) (see section 5.1.7) was successfully applied to determine the major ethnic groups in the world population and

to assign individuals correctly to one of these groups (Rosenberg et al., 2002; Bamshad et al., 2003; Tang et al., 2005). The assignment was most often corresponding to their self-reported ethnicity. Rosenberg et al. (2002) also tried to determine the structure within Europe based on 377 marker loci but only an average of 20 individuals per European population. The predefined populations could not be identified, the runs of STRUCTURE showed inconsistent results and, if at all, Basques and Sardinians could be distinguished from the other European populations. However, as already mentioned the sample size of European individuals was quite small. It still has to be empirically investigated how successful clustering approaches can be applied for European populations if the number of loci and the number of individuals is large enough. This aspect is later also addressed in our simulation study.

6.1.2 The impact of population stratification on case-control studies

There is an ongoing debate in the literature if unobserved population stratification is a serious problem for case-control association studies and if failures to replicate findings from case-control association studies are really attributable to population stratification. In this context we especially would like to mention the discussion between Thomas and Witte (2002) and Wacholder et al. (2002). Thomas and Witte (2002) raise serious concerns about the impact of population stratification whereas Wacholder et al. (2002) are of the opinion that population stratification is not a major threat for the validity of case-control studies. Thomas and Witte (2002) give some classical examples for population stratification in case-control studies, e.g. the study of Knowler et al. (1988) showed that a failure to adjust for population stratification would produce a spurious association between variants of the immunoglobulin gene *Gm* and type-2 diabetes in American Indians. However, this association was not causal and instead reflected confounding by the degree of Caucasian inheritance. Another example is that numerous studies which investigated the association between the $A1$-allele at the D_2 dopamine receptor locus and alcoholism gave contradictory results which might be explained by population stratification. However, Wacholder et al. (2002) pointed out that none of their examples is a demonstration of population stratification misleading the scientific community. Either population stratification could be corrected for by adjusting for ethnicity as in the first example or alternative explanations could be found for the failures to replicate positive findings as in the second example. Thus, in cases where ethnicity can be easily determined, population stratification should not be a major problem. A more serious concern is hidden population stratification. However, it is not at all clear that the large number of false positive findings in case-control studies really is attributable to population stratification. There are often many other possible reasons for false positive associations. One of the most important problems is poor epidemiological design especially with respect to control selection violating the basic design principles as for example explained in Rothman (1986). Further reasons are dif-

ferences in phenotype and case definition between the numerous studies or the impact of an unmeasured environmental confounder. Moreover, there are no associations between most alleles and specific diseases and truly positive genetic effects are small. Thus, a high proportion of false positives is inevitable when testing unlikely hypotheses especially with low power. It also has to be taken into account that usually multiple comparisons are carried out testing many genes and analyzing many subgroups. Often the analysis is not properly adjusted for multiple comparisons or only statistically significant results are reported leading to a publication bias.

To assess the impact of population stratification on the association results it has to be considered whether the circumstances leading to an increased number of false positive associations commonly exist. The first aspect is if the population heterogeneity in allele frequencies is large enough. There are some genes which are highly polymorphic and show large variations within and among populations. Genes controlling the immune response to infections, as for example HLA genes, belong to this category because populations have historically been subjected to different infections often killing people before they reach reproductive age. However, the majority of genes does not show such large variation and the genetic differences between populations are summarized in the F_{ST}-values which are rather small within central Europe (section 6.1.1).

The second aspect is which amount of heterogeneity in disease rates usually has to be expected. Cancer incidences, for example, are particularly well documented in Parkin et al. (2002). Within Europe, remarkable variation of cancer incidence rates can be observed for example for melanoma of skin in men. The annual age-standardized incidence rate per 100,000 ranges from 2.6 ± 0.16 in Lithuania to 14.3 ± 0.32 in Norway. Even the variation within a country can be large. For example, within England, rates vary from 4.8 ± 0.2 in Yorkshire to 8.9 ± 0.2 in the southern and western regions.

At first glance, the differences seem too small to cause major problems with population stratification. Indeed, the bias as calculated in proposition 3.5 is expected to be quite small but the variance inflation of the test statistic which is dependent on the sample size can be considerably high in large case-control studies. In the simulation chapter the variance inflation is theoretically calculated for different scenarios of confounding by population stratification. Most of these are realistic for European case-control studies (table 7.5).

Recently, new attempts were made to assess the impact of population stratification empirically based on real marker data. In the largest application up to date Freedman et al. (2004) investigated 11 case-control and case-cohort association studies by analyzing data from 24 to 48 unlinked single nucleotide polymorphisms in 90 to 500 cases and 69 to 500 controls. None of the studies showed significant evidence for stratification. However, confidence intervals for inflation factors where sufficiently broad that substantial levels of stratification could not be excluded. Increasing the number of markers to 114 single nucleotide polymorphisms significant evidence for stratification could be found in an African

American prostate cancer study. The inflation factor projected to a sample size of 1000 cases and 1000 controls was estimated at $\lambda = 1.5$. The observation of population stratification in this study is not entirely unexpected because African Americans of west African descent are thought to have a higher genetic risk than those with European descent. Thus, they could be overrepresented in the case sample.
A further investigation (Helgason et al., 2005) is based on the analysis of population structure in the Icelandic population which is expected to be rather homogeneous given its recent origin, small size and geographical isolation. The sample size is extremely large, 43,748 Icelanders divided into 3 birth cohorts and 11 geographical regions are genotyped at 40 microsatellite marker loci. The fixation index F_{ST} is estimated at 0.00338 in the $1895 - 1935$ cohort and falls to 0.00017 in the $1960 - 2000$ cohort indicating gradual admixture in the last generations. The impact of population stratification is assessed by simulating case-control status and randomly assigning individuals to cases and controls according to different sampling schemes from the geographical regions. Maximum values for λ in a sample of 1000 cases and 1000 controls were estimated at 1.24 for the $1895-1935$ cohort and 1.08 for the $1960 - 2000$ cohort. Thus, even in a rather homogeneous population as in Iceland there is notable regional subdivision possibly leading to a slight variance inflation.
Very recently, Campbell et al. (2005) presented the first example where stratification caused a spurious association in a sample of European Americans which was supposed to be homogeneous. However, the spurious association did not appear in a real case-control sample. Instead the sample was selected to demonstrate the impact of population stratification choosing height as a phenotype which which varies widely across Europe. The case-control sample comprised 1057 small and 1132 tall individuals. Estimation of the variance inflation factor in a subsample did not show any evidence of stratification. However, a marker with wider spread in allele frequencies among European populations that differ in hight might seem to be false positive associated. In fact, such a SNP in the gene *LCT* showed a strong association with height. The association was largely or completely due to stratification because it markedly decreased after stratifying the data according to grandparental ancestry and it could not be detected in two further studies. Thus, this example shows that markers which vary wider in frequency are more likely to produce spurious associations and these cannot be prevented by Genomic Control based on null loci with less frequency variation.

6.2 The German Genomic Control Study

Within the framework of the German National Genome Research Network (NGFN) a Genomic Control study was conducted to assess the impact of population stratification on case-control studies within Germany. The study is a joint effort of the Genetic Epidemiological Centers of Excellence (GEM = Genetisch epidemiologisches Methodenzentrum)

and the national genotyping platform. The study was mainly analyzed by the GEM in Bonn but we also contributed to the analysis by proposing and calculating the prediction rate as a new measure of predicting subpopulation membership (see section 2.4.2). As a joint work within the NGFN an article (Steffens et al.) already is submitted which contains the results of the NGFN Genomic Control study including the prediction rate. Here the description of the study and the main results are summarized briefly with an emphasis on the prediction rate, a more detailed description of the other issues can be found in Steffens et al..

6.2.1 The study sample

Population samples of unrelated individuals were recruited from three ongoing cross-sectional epidemiological surveys of regional German populations: KORA (Co-operative Health Research in the Region of Augsburg) from Southern Germany, POPGEN (Population Genetic Cohort) from Schleswig-Holstein in the north of Germany, and SHIP (Survey of Health in Pommerania) from Northeast Germany. Samples of more than 700 people from each survey (KORA: 730, POPGEN: 720, SHIP: 709) were genotyped at 212 SNP marker loci resulting in over 457,000 genotypes. The samples were matched for age (54 ± 13 years) and gender (50% males/ 50% females). All individuals were genotyped at the same 212 SNP marker loci, which were subdivided into three marker sets named according to the location of the genotyping center (GCKiel: 68 loci, GCMunich: 68 loci, GCBerlin: 76 loci + 3 duplicate loci). The set GCKiel is a set of coding SNPs located in functional genes and causing an amino acid exchange in the resulting protein. Thus, this marker set is potentially subject to selective forces. Under selective pressure allele frequencies are influenced by different probabilities of survival of individuals to reproductive age. In contrast, GCMunich and GCBerlin are two sets of neutral SNPs which are located far from known genes and are expected to be selectively neutral. Details of the marker selection are also given in Steffens et al.. The joined marker sets (GCBerlin, GCMunich) and (GCBerlin, GCKiel, GCMunich) will be referred to as "GC2BM" and "GC3BKM", respectively.

6.2.2 Population structure in the study sample

Table 6.1 shows the F_{ST}-values estimated from the different marker sets by the GEM in Bonn. In the majority of cases, F_{ST}-estimates were positive but quite low taking values from -0.0002 to 0.0008. The highest F_{ST}-estimates ranging from 0.0003 to 0.0008 for the different marker sets were identified for the geographically most distant populations KORA from Southern Germany and SHIP from Northeast Germany. F_{ST}-estimates between KORA and POPGEN, the second population from North Germany, were lower

	GCKiel	GCMunich	GCBerlin
KORA, POPGEN	0.0000 ± 0.0001	0.0001 ± 0.0001	0.0004 ± 0.0002
KORA, SHIP	0.0003 ± 0.0002	0.0008 ± 0.0003	0.0005 ± 0.0002
POPGEN, SHIP	-0.0002 ± 0.0001	0.0001 ± 0.0001	0.0001 ± 0.0002
TOTAL	0.0001 ± 0.0001	0.0004 ± 0.0002	0.0003 ± 0.0001
	GC2BM	GC3BKM	
KORA, POPGEN	0.0003 ± 0.0001	0.0002 ± 0.0001	
KORA, SHIP	0.0007 ± 0.0002	0.0005 ± 0.0001	
POPGEN, SHIP	0.0001 ± 0.0001	0.0000 ± 0.0001	
TOTAL	0.0003 ± 0.0001	0.0002 ± 0.0001	

Table 6.1. Estimated F_{ST} ± standard error in the Genomic Control Study, calculated by the GEM Bonn. F_{ST} is estimated according to the formulas described in Weir and Cockerham (1984); Weir and Hill (2002).

varying from 0 to 0.0004. However, these estimates are at least partly significantly different from 0 if an underlying normal distribution is assumed for calculating confidence intervals from estimators and standard errors. Between the two populations from North Germany POPGEN and SHIP no significant difference could be observed, F_{ST}-estimates vary around zero and one of the F_{ST}-estimates is even negative. The comparison between coding SNPs (GCKiel) and non-coding SNPs (GCBerlin, GCMunich) shows that F_{ST}-estimates are higher for non-coding SNPs. The F_{ST}-value of 0.0007 estimated from non-coding SNPs between the most distant populations KORA and SHIP is $1.5 - 3$ times lower than the F_{ST}-value between Germans and other Germanic populations (Dutch, Danish, English, Austrian, Swiss, Belgians) (see section 6.1.1).
The inbreeding coefficients F_{IS} were also calculated by the GEM Bonn. These values were approximately one order higher than the F_{ST}-values. For the coding SNPs F_{IS} was estimated at 0.0038 averaged over all populations whereas for the non-coding SNPs an estimate of 0.0074 was obtained. Thus, the Genomic Control study again confirms the well-known fact that most variability in human populations is observed within populations and only a minor fraction of genetic variation due to differences between populations. This again shows that the assumption of Hardy-Weinberg equilibrium within the subpopulations which is used in the clustering algorithm is quite idealistic.
To evaluate if the prediction of the population membership is possible based on the given marker loci we calculated the prediction rate introduced in section 2.4.2. In most cases the prediction rate was only slightly larger than 50% comparing any two population samples with slight advantages for the marker sets including null loci (GCBerlin, GCMunich) in opposite to the marker set including coding SNPs (GCKiel) (see table 6.2). Comparing "KORA versus SHIP" it is slightly growing with an increasing number of marker loci up to 54% in the whole data set GC3BKM whereas for "KORA versus POPGEN" this tendency if at all existent is even minor. For "POPGEN versus SHIP" the prediction rate is always

	GCKiel	GCMunich	GCBerlin
KORA, POPGEN	0.5005 [0.4926, 0.5084]	0.5027 [0.4961, 0.5094]	0.5119 [0.5028, 0.5211]
KORA, SHIP	0.5101 [0.5007, 0.5196]	0.5203 [0.5055, 0.5350]	0.5174 [0.5076, 0.5273]
POPGEN, SHIP	0.4952 [0.4898, 0.5007]	0.5019 [0.4958, 0.5081]	0.5076 [0.4979, 0.5173]
TOTAL	0.3351 [0.3300, 0.3401]	0.3406 [0.3339, 0.3473]	0.3459 [0.3386, 0.3532]
	GC2BM	GC3BKM	
KORA, POPGEN	0.5143 [0.5030, 0.5256]	0.5139 [0.5014, 0.5265]	
KORA, SHIP	0.5353 [0.5197, 0.5510]	0.5421 [0.5261, 0.5580]	
POPGEN, SHIP	0.5090 [0.4981, 0.5199]	0.5046 [0.4930, 0.5162]	
TOTAL	0.3518 [0.3426, 0.3609]	0.3526 [0.3430, 0.3621]	

Table 6.2. Prediction rate in the Genomic Control Study, calculated as described in section 2.4.2.

around 50% independently of the marker set and the confidence interval shows that there is no significant difference between POPGEN and SHIP. Here, the estimate is even less than 50% using the GCKiel marker set what is conform with the negative estimate for F_{ST} in this case. Thus, the results for the prediction rate confirm the trend in the F_{ST}-values that there is some small population difference between North and South Germans but no difference within North Germany. However, even between the most different populations KORA and SHIP the prediction rate is very low using the total marker set of more than 200 markers.

Given the low prediction rate it is not surprising that different clustering algorithms failed to detect any population structure in the sample. We applied the standard EM algorithm to the data but failed to detect any STRUCTURE in the sample. The highest AIC was achieved for $K = 1$ and the runs for $K = 2$ lead to very inconsistent results depending on the starting values. This indicates that there is no obvious population structure. The group in Bonn also applied the MCMC algorithm implemented in the program STRUCTURE (see section 5.1.7) to the data set. Different admixture models and prior distributions were used. However, the MCMC algorithm also failed to detect any structure in the complete sample. The insensitivity of these algorithms to detect small structure is also concordant with our simulation results for the EM algorithm (see section 7.2.1). In our simulations we show that based on 100 SNPs population stratification generally could be detected for $F_{ST} = 0.0050$ but not for $F_{ST} = 0.0025$ anymore using an EM algorithm as clustering method. Since the Bayesian model implemented in STRUCTURE is also a probability based clustering method a similar performance is expected.

6.2.3 Consequences for case-control studies within Germany

The main research question associated with the Genomic Control study is to assess the impact of population stratification on case-control studies within Germany and the possi-

	GCKiel	GCMunich	GCBerlin
KORA, POPGEN	0.968 [0.710, 1.398]	1.057 [0.769, 1.545]	1.433 [1.060, 2.047]
KORA, SHIP	1.379 [1.012, 1.991]	1.875 [1.370, 2.724]	1.816 [1.317, 2.663]
POPGEN, SHIP	0.737 [0.540, 1.064]	1.016 [0.739, 1.486]	1.404 [1.018, 2.059]
	GC2BM	GC3BKM	
KORA, POPGEN	1.256 [1.004, 1.618]	1.160 [0.964, 1.423]	
KORA, SHIP	1.846 [1.467, 2.395]	1.685 [1.396, 2.074]	
POPGEN, SHIP	1.208 [0.959, 1.571]	1.044 [0.864, 1.287]	

Table 6.3. Inflation factor in the Genomic Control Study in the worst case scenario. The mean-based estimator for the inflation factor of the allelic χ^2-test is calculated as proposed by Reich and Goldstein (2001) (see section 4.2.1) and the confidence interval as given in proposition 4.4.

bilities to correct for stratification. Since clustering algorithms do not detect any structure, methods of Structured Association seem to be inappropriate to correct for any confounding effects. However, it has to be investigated if the clustering process could be more successful when increasing the number of null loci.
To assess whether there could be nevertheless considerable variance inflation for a case-control study within Germany we estimate the variance inflation factor for the worst-case-scenario that cases are recruited from one population and controls are recruited from another of the three populations (see table 6.3). The inflation factors again show the same northern-southern trend as before. Although F_{ST}-estimates are rather small the estimated variance inflation is considerably large comparing KORA and SHIP ranging from 1.379 for the GCKiel set to 1.875 for the GCMunich set. However, if cases and controls are recruited in a similar sampling scheme from different German regions, then only a very minor inflation is expected to be observed if the formula for the variance inflation is considered (see proposition 4.2). In this case, population structure within Germany should not be a major problem.

6.3 Discussion

Summarizing the recently published applications the opinion seems to prevail that even in populations that seem to be rather homogeneous there can be a measurable impact of hidden population stratification on the association results. This is also the result from analyzing the Genomic Control study. Within Germany methods of Structured Association cannot be successfully applied for correcting case-control association tests, but there could be a measurable variance inflation when recruiting cases and controls from different geographical areas. Within some parts of Europe the stratification is larger and this amount of stratification is investigated later in the simulations.
The results also suggest to take care when designing a study. The examples show that recruiting cases and controls from different geographical areas within a country should be

avoided. Moreover, care should be taken in the analysis if study participants are collected from different countries. Even if no additional markers are genotyped, the analysis could be stratified by the country of origin to avoid huge levels of population stratification.

7 Simulations

In our simulations we intend to simulate the realistic situation of large association studies with moderate population stratification as in Europe. As in Köhler and Bickeböller (2006) we want to investigate for different simulation scenarios which statistical method is most appropriate to correct for population stratification when testing for association. On the one hand the two approaches Genomic Control and Structured Association shall be compared in general and on the other hand the differences between the Structured Association approaches including our own development shall be investigated. For Structured Association only the two-step approach is simulated applying the standard or the phenotype-dependent EM algorithm for clustering. The reason is that we want to simulate a sufficient number of multilocus data sets and the candidate locus with an adequate number of replications for each of these multilocus data sets. This cannot be implemented using a Bayesian clustering approach as Pritchard et al. (2000a) or a one-step approach as Satten et al. (2001) because of time constraints.

7.1 The set-up of the simulation study

7.1.1 The simulation model

We simulate multilocus marker data for K discrete subpopulations. For each locus we draw an ancestral allele frequency φ_l from a uniform distribution in (0.1,0.9) to avoid very rare alleles. Subpopulation allele frequencies are generated in the beta-binomial model (see section 2.3.2). The two alleles of an individual from subpopulation k are independently drawn from a Bernoulli distribution with subpopulation allele frequency φ_{kl}. Thus, for the simulations we also use a discrete subpopulation model just as we assumed for the inference on population structure with the EM algorithm.
Ancestral allele frequencies at the candidate locus are again randomly chosen from a uniform distribution in (0.1,0.9). The candidate locus is then simulated in a multiplicative penetrance model (see section 3.2.2). For all subpopulations $k = 1, \ldots, K$ the allelic relative risks ψ_k are specified. The allele frequencies $p_k^{(u)}$ are directly simulated in the beta-binomial model since the control group approximately represents the original population if the disease prevalence is low. Allele frequencies among cases are calculated as derived in proposition 3.1 applied for each subpopulation k separately under the assumption that the allele frequencies in the original population are equal to the allele frequencies within the control group. To generate the multilocus marker data the beta-binomial model is employed with different parameter configurations as described below.

7.1.2 Details of the simulations

In the simulations the AIC is applied for choosing the number of subpopulations K (see section 5.1.5). A problem is that the EM algorithm converges to different local maxima leading to different values of the log-likelihood dependent on the starting values. Thus, we run the EM algorithm with different starting values for the same K. Since we need a well defined criterion for our simulations to choose a specific run of the EM algorithm to calculate AIC(K) we take the run with the median log-likelihood of all runs for the same K. This prevents us from taking a global maximum which potentially lies close to the boundary of the parameter space and is rarely reached. In all unambiguous cases where the global maximum is reached in more than 50% of all runs, the median log-likelihood equals the log-likelihood for the global maximum. A further problem is that it may happen that the AIC reaches a local minimum for a small number of subpopulations first before it reaches its global minimum for a larger number. Here, we concentrate on the first local minimum rather than the global minimum, since an increase of the AIC after the first local minimum is an indication against an additional subpopulation.
After applying the EM algorithm we have to investigate how accurate the EM algorithm infers population structure for the different parameter configurations. Therefore, we estimate F_{ST} from the inferred subpopulations and compare it to the true simulated F_{ST}. Here the formula for estimating F_{ST} derived in section 2.4.1 is applied replacing the unknown number of individuals N_k in subpopulation k with the estimated number $\sum_{i=1}^{N} q_{ik}$. However, this is only an approximation since the formula is derived assuming subpopulation membership is known in advance. This leads to an underestimation of the true variance of $\widehat{\varphi}_{kl}$ in the case of unknown population structure and an overestimation of F_{ST} for small sample sizes N or few loci L. The true variances and covariances are difficult to determine because only parameter estimates without their variances are obtained as a result of the EM algorithm.
The main purpose of the simulations is to assess type-I error rate and power of the different association tests. Three SA test statistics *STRAT* and *P-STRAT* and *P-CHIPOP* are examined. *STRAT* and *P-STRAT* are based on the likelihood ratio test (see section 5.2.2) and *P-CHIPOP* on the Wald test (see section 5.2.3). For *STRAT* the standard EM has been applied to infer population structure whereas the P indicates that the phenotype-dependent EM algorithm has been used. For all three test statistics the p-value is based on the asymptotic distribution. Thus, although a simpler clustering algorithm is applied *STRAT* corresponds to the idea of Pritchard et al. (2000a,b) who propose clustering without considering phenotype information and then applying the likelihood ratio test as an association test. The standard EM is always applied for the same number of subpopulations which is inferred for the P-EM algorithm to make the *STRAT* and *P-STRAT* results more comparable by using the same number of subpopulations for both algorithms. The SA test statistics are compared to the simple χ^2-test *CHISQ* as well as to the two GC test

label	explanation	section
CHISQ	allelic χ^2-test	3.2.4
GC-MED	GC applied for Armitage's trend test, median-based estimator for λ	4.2.1
GC-MEAN	GC applied for the allelic χ^2-test, mean-based estimator for λ	4.2.1
STRAT	standard EM algorithm & likelihood ratio test	5.2.2
P-STRAT	phenotype-dependent EM algorithm & likelihood ratio test	5.2.2
P-CHIPOP	phenotype-dependent EM algorithm & Wald test	5.2.3

Table 7.1. Overview over the simulated tests for association in a stratified population

statistics *GC-MEAN* and *GC-MED*. Table 7.1 gives an overview over the simulated tests and the sections where these tests are described. Type-I-error rate, power in a homogeneous model of equal allelic relative risks and power in a heterogeneous model of different allelic relative risks in the subpopulations are investigated. As described before in a homogeneous model population stratification only acts as a confounder and not as an effect modifier in contrast to the heterogeneous model (see section 3.3.2). For each parameter configuration 100 sets of multilocus marker data were simulated. To each of these multilocus marker data sets the candidate locus was simulated with 10000 replications under the null hypothesis and the two alternatives. This approach has the advantage that not only the median or mean type-I error rate or power over the 100 multilocus marker sets can be determined but also the variation over different marker sets can be investigated.

7.1.3 The simulation parameters

For our simulations we first fix a basic parameter configuration which models a large association study with moderate population stratification. We simulate a population of $N = 2000$ individuals which consists of two discrete subpopulations. $L = 100$ loci are used to infer population structure. The distance between the subpopulations is fixed at $F_{ST} = 0.01$. This corresponds to a population structure which is expected between some pairs of central European populations which are more distantly related (see section 6.1.1). For example, the distance between Spanish and Swedish is estimated as $F_{ST} = 0.0099$ and the distance between Scottish and Czech as $F_{ST} = 0.0104$.
The simulated sample consists of an equal number of cases and controls. The prevalence is two-fold higher in subpopulation 2 than in subpopulation 1 which can be expressed as $RR = 2$ if subpopulation 2 is compared to subpopulation 1. Thus, the relative risk RR refers to the ratio of the general disease risks in two different subpopulations and is not related to any genetic effect. In the original population both subpopulations shall be equally represented. Hence if the prevalence is low the control group is approximately expected to have equal proportions of both subpopulations, i.e. $\pi_1^{(u)} = 0.5$ and $\pi_2^{(u)} = 0.5$. For the cases the given disease prevalence ratio leads to $\pi_1^{(a)} = \frac{1}{3}$ and $\pi_2^{(a)} = \frac{2}{3}$. Thus,

L	F_{ST}	N	K, RR
25	0.0025	500	1, -
50	0.005	1000	2, 1
100	**0.01**	**2000**	**2, 2**
200	0.02	4000	2, 4
400	0.04	8000	4, 2:3:4

Table 7.2. Variation of the basic parameter configuration in the simulation study. The basic parameter configuration is shown in bold. Each possible parameter configuration is obtained by substituting only one entry of the basic parameter set by a different entry of the corresponding column. Subpopulation 1 is always chosen as a reference for calculating the RR, hence for $K = 4$ three relative risks have to be specified.

altogether subpopulation 2 is slightly overrepresented with $\pi_1 = 0.417$ and $\pi_2 = 0.583$. For the simulations the number of cases and controls from each subpopulation are fixed to their expected values and rounded if necessary. Table 7.2 summarizes the characteristics of the basic parameter set and shows how it is varied later to investigate the influence of different parameters. The basic configuration is always varied in one parameter only except for K and RR which are jointly changed. The number of loci L ranges from 25 to 400, the fixation index F_{ST} from 0.0025 to 0.04, the number of individuals N from 500 to 8000 and the number of subpopulations K from 1 to 4 with different relative risks for $K = 2$. Altogether $17 = 1 + 4 \times 4$ parameter configurations are simulated.
For the power simulation in a homogeneous model a fixed alternative with an allelic relative risk of $\psi_k = 1.3$ is chosen leading to a power of approximately 80% for the χ^2-test and the basic parameter configuration. In the heterogeneous model the candidate gene is associated with the disease only in subpopulation 2 for all parameter configurations with $K = 2$. An allelic relative risk of $\psi_1 = 1.0$ and $\psi_2 = 1.5$ is chosen. For $K = 4$ the allelic relative risks of the two further subpopulations are selected in between, resulting in the vector of allelic relative risks $\boldsymbol{\psi} = (1.000, 1.167, 1.333, 1.500)$.

7.2 Results

7.2.1 Inference on population structure

When applying the P-EM algorithm the correct number of subpopulations is inferred for almost all data sets (97 out of 100) of the basic parameter configuration and the median F_{ST} is quite accurately estimated at 0.0104 (table 7.3). The identification rate which measures the average posterior probability of correct assignment of the individuals to the subpopulations is estimated giving the median at 78.3%. Although the general structure of the population is correctly detected it could not be reliably determined which

	inferred K							inferred F_{ST}			identification rate %		
	1	2	3	4	5	6	7	median	$q_{10\%}$	$q_{90\%}$	median	$q_{10\%}$	$q_{90\%}$
basic*	0	**97**	3	0	0	0	0	0.0104	0.0089	0.0122	78.3	75.2	80.9
L=25	7	**83**	9	1	0	0	0	0.0136	0.0103	0.0196	61.4	56.6	65.3
L=50	0	**84**	7	4	5	0	0	0.0117	0.0090	0.0160	68.5	63.7	71.6
L=100*	0	**97**	3	0	0	0	0	0.0104	0.0089	0.0122	78.3	75.2	80.9
L=200	0	**100**	0	0	0	0	0	0.0100	0.0089	0.0115	88.7	86.7	90.7
L=400	0	**100**	0	0	0	0	0	0.0100	0.0092	0.0109	96.7	96.0	97.5
F=0.0025	73	**24**	2	1	0	0	0	-	-	-	-	-	-
F=0.005	0	**89**	7	4	0	0	0	0.0064	0.0053	0.0079	67.0	61.8	70.9
F=0.01*	0	**97**	3	0	0	0	0	0.0104	0.0089	0.0122	78.3	75.2	80.9
F=0.02	0	**100**	0	0	0	0	0	0.0197	0.0168	0.0238	88.9	86.5	91.8
F=0.04	0	**98**	1	1	0	0	0	0.0388	0.0318	0.0463	96.9	95.0	98.1
N=500	3	**97**	0	0	0	0	0	0.0122	0.0106	0.0136	77.3	71.4	80.5
N=1000	0	**99**	1	0	0	0	0	0.0108	0.0090	0.0126	77.9	74.1	81.1
N=2000*	0	**97**	3	0	0	0	0	0.0104	0.0089	0.0122	78.3	75.2	80.9
N=4000	0	**98**	2	0	0	0	0	0.0102	0.0087	0.0122	78.4	75.8	81.2
N=8000	0	**99**	1	0	0	0	0	0.0101	0.0083	0.0120	78.5	75.3	81.5
K=1	**96**	4	0	0	0	0	0	-	-	-	-	-	-
K=2,RR=1	0	**100**	0	0	0	0	0	0.0101	0.0086	0.0123	77.0	74.0	80.6
K=2,RR=2*	0	**97**	3	0	0	0	0	0.0104	0.0089	0.0122	78.3	75.2	80.9
K=2,RR=4	0	**92**	8	0	0	0	0	0.0101	0.0089	0.0124	80.3	77.6	82.5
K=4	0	0	12	**57**	20	10	1	0.0128	0.0105	0.0196	57.4	52.5	60.5

Table 7.3. Structure inferred by the P-EM algorithm for the 100 multilocus data sets simulated for each parameter configuration. The basic parameter configuration is denoted by * and consecutively varied in L, F_{ST}, N and finally K and RR simultaneously. It is only simulated once but shown in the first row and in the series for each parameter in turn. The number of data sets where K is correctly inferred is shown in bold. F_{ST} is estimated for all data sets where the inferred K is larger than 1. The *identification rate* measures the concordance of the inferred and the simulated population structure. It is defined as mean posterior probability of an individual of being assigned to the correct subpopulation, averaged over all individuals. The identification rate is only calculated if the correct number of subpopulations is inferred. The F_{ST}-values and identification rates shown here are the median, 10%-quantile and 90%-quantile taken over the included multilocus data sets.

		simulated proportion of S_2	estimated proportion of S_2		
			median	$q_{10\%}$	$q_{90\%}$
complete sample	phenotype-dependent EM	0.583	0.577	0.538	0.615
	standard EM	0.583	0.575	0.536	0.617
cases	phenotype-dependent EM	0.667	0.660	0.614	0.698
	standard EM	0.667	0.622	0.568	0.660
controls	phenotype-dependent EM	0.500	0.493	0.453	0.539
	standard EM	0.500	0.530	0.487	0.576
difference cases - controls	phenotype-dependent EM	0.167	0.163	0.137	0.191
	standard EM	0.167	0.089	0.071	0.108

Table 7.4. Simulated and estimated proportion of subpopulation 2 (S_2) within cases and controls for the standard EM in comparison to the phenotype-dependent EM (P-EM) calculated for the basic parameter configuration. The estimates shown here for the subpopulation proportions and the difference are the median, 10%-quantile and 90%-quantile over the 97 multilocus data sets where the correct number of subpopulations is inferred.

individual belongs to which subpopulation. Interestingly, even if the number of loci is decreased down to 25, in most cases it is possible to identify the two subpopulations. Varying the number of loci from 25 to 400 the whole range from a very poor identification of the subpopulations (61.4%) to a nearly perfect identification (96.7%) is observed. The variation over F_{ST} has a similar effect on the clustering results as varying the number of loci. If F_{ST} is equal to 0.0025 the identification of the two subpopulations is not anymore possible with 100 loci. Variation of the sample size N does not have a large effect on the identification rate, the median identification rate is always between 77% and 79% in the considered sample size range. Thus, an increased sample size does not help to infer population structure. The estimate of F_{ST} is substantially biased for smaller sample sizes N, especially $N = 500$, due to the estimation method (see section 7.1.2). Our simulations include different numbers of subpopulations starting from $K = 1$ where no structure is present. The sample is correctly identified as homogeneous in most cases. For $K = 2$ the clustering results are only weakly dependent on the RR and hence on the subpopulation proportions. For $K = 4$ it is more difficult to infer the correct number of subpopulations and the identification rate is much lower than for the basic parameter configuration (57.4%).

For the basic parameter configuration we also evaluate the deviation of the estimated subpopulation proportions from the simulated proportions within cases and controls when applying the standard EM in comparison to the P-EM algorithm (table 7.4). With the P-EM algorithm the subpopulation proportions are estimated quite accurately for the whole sample, as well as separately for cases and controls. The standard EM algorithm shows the expected deviation. Here, a quite accurate estimation of the subpopulation proportions is possible for the whole sample, but within cases the true proportion of

subpopulation 2 is underestimated and within controls overestimated. The median difference of the proportion of subpopulation 2 between cases and controls is 0.089 instead of 0.167 which corresponds to only 53.3% of the true difference and even the 90%-quantile of 0.108 is clearly smaller than the simulated difference.

7.2.2 Theoretical variance inflation for the simulation scenarios

Considering the association tests, first some theoretical properties of Genomic Control are calculated to theoretically assess the impact of population stratification for the different parameter configurations (table 7.5). The basic data set corresponds to a theoretical variance inflation λ of 1.55 calculated from $N^{(a)} = N^{(u)} = 1000$, $F_{ST} = 0.01$ and the factor $\frac{1}{18}$ which depends on the subpopulation proportions according to proposition 4.2. Varying the various parameters inflation factors between 0.99 and 3.21 are obtained. Since $\lambda - 1$ is approximately proportional to N for an equal number of cases and controls and also to F_{ST} (proposition 4.2), $\lambda - 1$ is approximately sixteen times higher for $N = 8000$ than for $N = 500$ and also for $F_{ST} = 0.04$ in comparison to $F_{ST} = 0.0025$. The length of the interval where the mean-based estimator of λ lies in with 80% probability is 0.56 for the basic configuration and depends on the number of loci as well as on the variance inflation itself.
The expected type-I error rate of the χ^2-test for a nominal level of 0.01 is 0.0384 for the basic parameter configuration. Thus, the hypothesis is almost 4 times more often rejected than it is supposed to be. The expected type-I error rate only depends on the variance inflation λ and varies from 0.0096 to 0.1508. In contrast, the 80%-interval for the observed type-I error rate of *GC-MEAN* is not dependent on the variance inflation λ as long as always the unbounded estimator $\widehat{\lambda}$ is used. It only depends on the number of loci L. For the basic parameter set and all other configurations with 100 loci the interval is [0.0050, 0.0194]. Hence, the 90%-quantile for the observed type-I error rate is 1.94 times larger than the nominal level. Thus, even applying GC there is a remaining probability of 10% to choose a set of null loci leading to a rejection of the null hypothesis at least 1.94 times more often than allowed when testing several non-associated candidate loci using that set of null loci.

7.2.3 Type-I-error rate of the association tests

The main focus of our simulation study is on comparing type-I error rate and power of the different association tests. Figure 7.1 shows the simulated type-I error rates for a nominal level of 0.01. The type-I error rates of the χ^2-statistic are close to their expected values calculated in table 7.5. This is also true for the highest rates which are not covered by the

	variance inflation			type-I error rate		
	true λ	$\widehat{\lambda}$ (mean-based)		χ^2-test	*GC-MEAN*	
		$q_{10\%}$	$q_{90\%}$		$q_{10\%}$	$q_{90\%}$
basic*	1.55	1.27	1.83	0.0384	0.0050	0.0194
L=25	1.55	1.02	2.13	0.0384	0.0025	0.0365
L=50	1.55	1.17	1.96	0.0384	0.0038	0.0253
L=100*	1.55	1.27	1.83	0.0384	0.0050	0.0194
L=200	1.55	1.35	1.75	0.0384	0.0062	0.0160
L=400	1.55	1.41	1.69	0.0384	0.0071	0.0140
F=0.0025	1.14	0.94	1.35	0.0157	0.0050	0.0194
F=0.005	1.27	1.05	1.51	0.0225	0.0050	0.0194
F=0.01*	1.55	1.27	1.83	0.0384	0.0050	0.0194
F=0.02	2.10	1.73	2.48	0.0752	0.0050	0.0194
F=0.04	3.19	2.63	3.78	0.1493	0.0050	0.0194
N=500	1.13	0.93	1.34	0.0154	0.0050	0.0194
N=1000	1.27	1.04	1.50	0.0220	0.0050	0.0194
N=2000*	1.55	1.27	1.83	0.0384	0.0050	0.0194
N=4000	2.10	1.73	2.49	0.0754	0.0050	0.0194
N=8000	3.21	2.65	3.81	0.1508	0.0050	0.0194
K=1	1.00	0.82	1.18	0.0100	0.0050	0.0194
K=2,RR=1	0.99	0.82	1.17	0.0096	0.0050	0.0194
K=2,RR=2*	1.55	1.27	1.83	0.0384	0.0050	0.0194
K=2,RR=4	2.79	2.30	3.31	0.1230	0.0050	0.0194
K=4	1.49	1.23	1.77	0.0348	0.0050	0.0194

Table 7.5. Theoretical variance inflation for the different parameter configurations and the theoretical influence of the estimation on the type-I error rate of the test statistic *GC-MEAN*. The theoretical variance inflation λ is calculated as derived in proposition 4.2 and the interval where the mean-based estimator lies in with 80% probability according to proposition 4.4. The expected type-I error rates are calculated for a nominal level of 0.01. Here the formulas given in proposition 4.5 are applied to determine the expected type-I error rate for the χ^2-test and the interval where the type-I error rate of *GC-MEAN* lies in with 80% probability if always the unbounded estimator for λ is used.

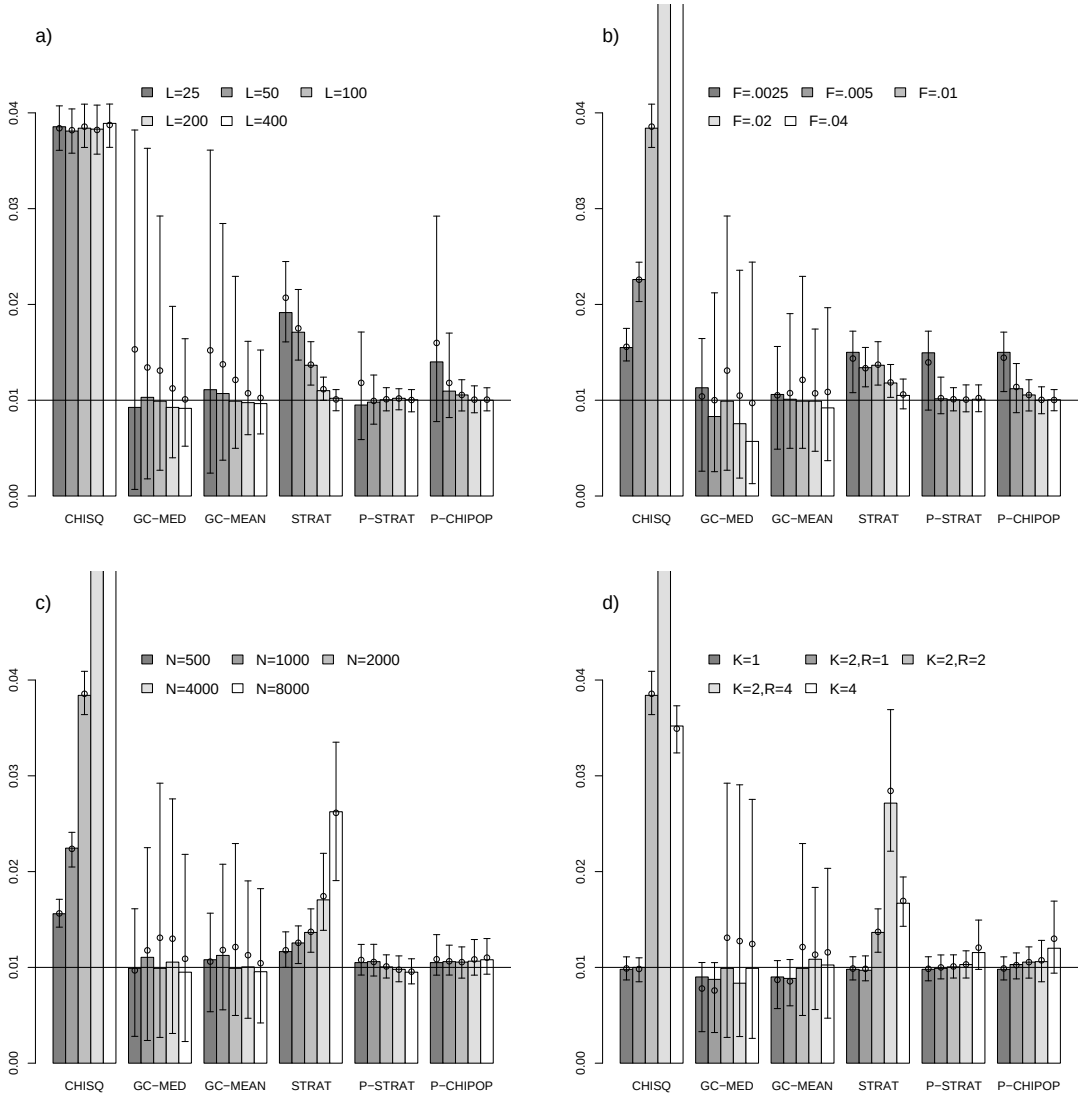

Figure 7.1. Simulated type-I error rates for a nominal level of 0.01. The four graphics show the influence of a) L, b) F_{ST}, c) N and d) K and RR on the type-I error rate. The bars show the median over the 100 multilocus data sets, the points indicate the mean and the intervals range from the empirical 10%- to the 90%-quantile.

range of the y-axis. *GC-MEAN* also shows the expected behaviour. The median type-I error rate is always close to the nominal level of 0.01 and the empirical quantiles agree in principle with their theoretical quantiles from table 7.5, as much as possible with only 100 simulations of multilocus data. For small inflation factors it should be noted that the 90%-quantile is bounded by the type-I error rate of the χ^2-test since GC is applied with a lower bound of $\widehat{\lambda} = 1$. This yields GC to be conservative for small inflation factors. *GC-MED* shows a similar pattern with even larger variation over the 100 multilocus data sets. This can be explained by the larger variance of the median in comparison to the mean. Interestingly, for large population structure *GC-MED* seems to be conservative for the median of the 100 data sets but not for the mean. As Marchini et al. (2004) we also observe that both GC statistics are somewhat inflated for a small number of loci when considering the mean over the 100 data sets instead of the median.
The comparison of the three SA tests reveals that *STRAT* is highly inflated for some parameter sets whereas the other two statistics *P-STRAT* and *P-CHIPOP* based on the P-EM algorithm have the correct type-I error rate in most cases. The inflation of *STRAT* especially increases with high RR or a large number of individuals. In the first situation the bias introduced by the standard EM algorithm is expected to be larger because the subpopulation proportions within cases and controls are more different from each other and in the second situation the absolute amount of misclassification due to the wrong model increases. *P-STRAT* and *P-CHIPOP* have the correct median type-I error rate in all situations where at least the general population structure could be inferred correctly. Moreover, the variation over the 100 multilocus marker sets is small. Even in situations where the correct assignment of individuals to the subpopulations is rarely possible as shown in table 7.3, e.g. $L = 50$ or $F = 0.005$, *P-STRAT* and *P-CHIPOP* are quite accurate. Only if the number of loci L or F_{ST} decrease further and clustering is not anymore possible, they are clearly inflated with *P-STRAT* being less affected than *P-CHIPOP*. An inflation is also visible if the number of subpopulations increases to $K = 4$ because the inference on population structure then becomes more difficult.
Altogether the type-I error rate simulation shows that both the SA test statistics based on the P-EM algorithm and GC maintain the correct type-I error rate with some exceptions, but the variation over the 100 multilocus marker sets is much lower applying SA. Since the variation of the type-I error rate of GC is not dependent on F_{ST} in contrast to SA, for increasing F_{ST} SA becomes even more superior.

7.2.4 Power of the association tests

Power simulations in a homogeneous model of equal allelic relative risks in all subpopulations are shown in figure 7.2. In the homogeneous model *P-CHIPOP* generally has the highest power of all test statistics, even a slightly higher power than the highly inflated χ^2-test. The power of *P-STRAT* is lower and the power difference between *P-STRAT*

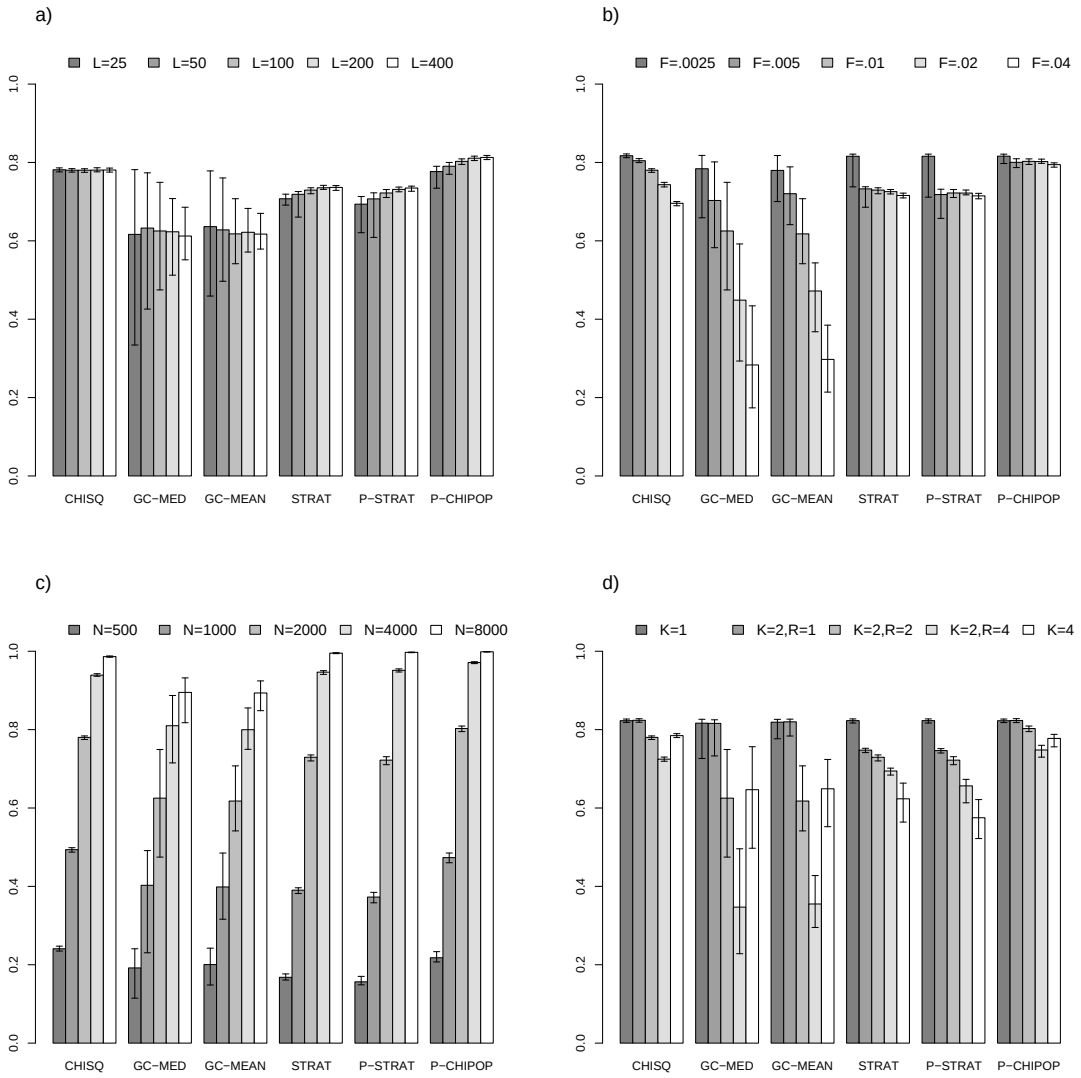

Figure 7.2. Power simulations in a homogeneous model of equal allelic relative risks in all subpopulations. The four graphics show the influence of a) L, b) F_{ST}, c) N and d) K and RR on the power. A fixed alternative with an allelic relative risk of $\psi_k = 1.3$ is chosen. The power is also calculated for a nominal level of 0.01, i.e. the same theoretical cut-off-values as for the type-I error rate simulation are used. The power of the allelic χ^2-test and *STRAT* is shown for comparison although the type-I error rate is highly inflated. The bars show the median over the 100 multilocus data sets and the intervals range from the empirical 10%- to the 90%-quantile.

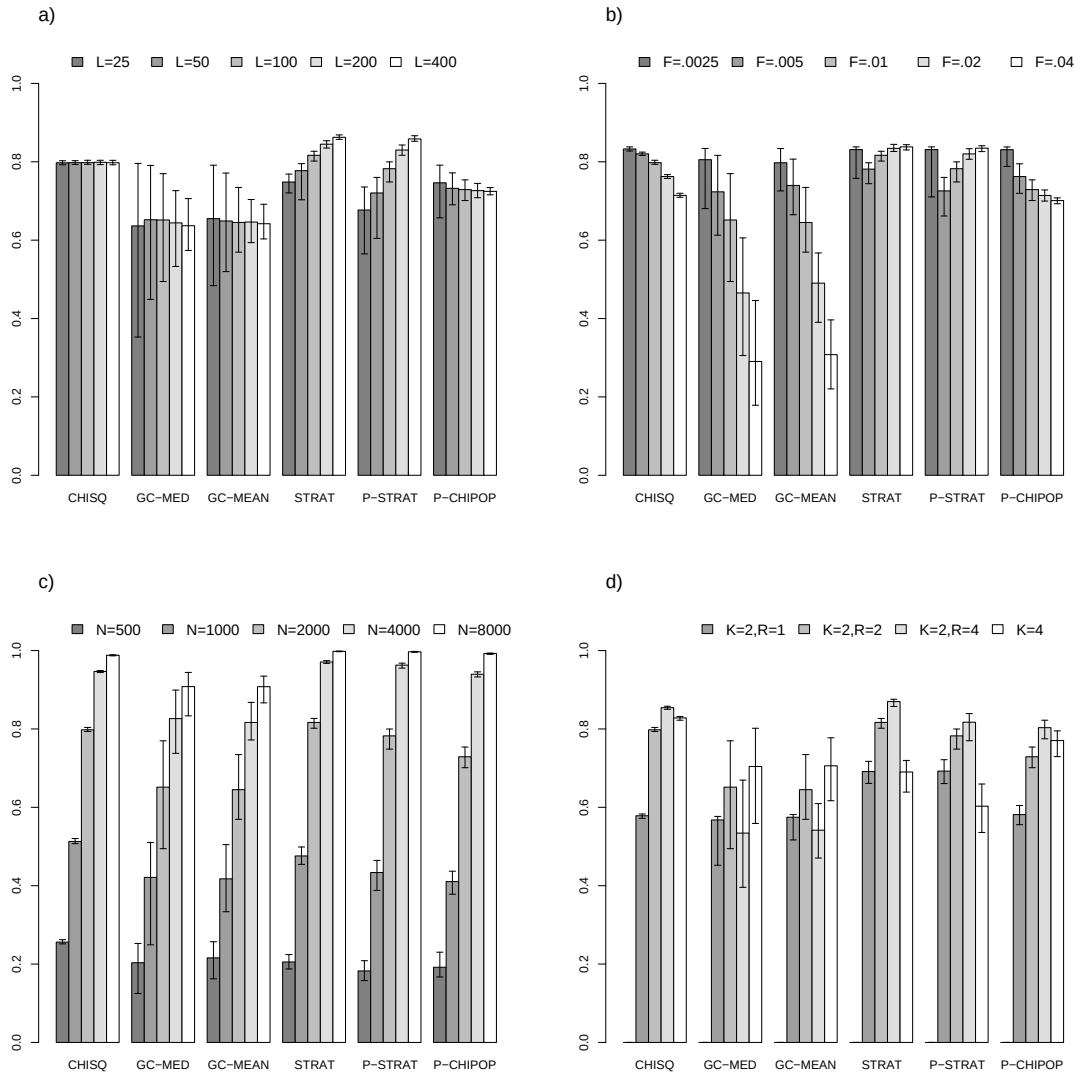

Figure 7.3. Power simulations in a heterogeneous model of different allelic relative risks in all subpopulations. The four graphics show the influence of a) L, b) F_{ST}, c) N and d) K and RR on the power. Details of the simulated alternative are described in section 7.1.3. Further details of the power simulation are the same as for the homogeneous model (see figure 7.2).

and *P-CHIPOP* is approximately 8% for absolute power values around 70% to 80% for $N = 2000$. An exception is the case $K = 4$. Here the power of *P-STRAT* is drastically lower than the power of *P-CHIPOP* with a difference of 20.2% in the medians. The additional loss of power due to the increasing number of subpopulations can be explained by a comparison of the degrees of freedom of the two test statistics. *P-CHIPOP* always has 1 df whereas for *P-STRAT* the degrees of freedom correspond to the number of subpopulations K. In most situations the GC test statistics have an even lower power than all SA test statistics and show again higher variation. The power is rapidly decreasing for a high F_{ST} and large population structure as well as for a high RR and hence different population structure within cases and controls. This is plausible because a model based approach should be superior to simply estimating an inflation factor, especially in these situations. In some cases, including small sample size ($N = 500$), equal subpopulation structure within cases and controls ($RR = 1$) and more than two subpopulations ($K = 4$) GC is slightly superior to *P-STRAT* but not to *P-CHIPOP*.
The power simulations in a heterogeneous model of different allelic relative risks in the subpopulations give a somewhat different result (figure 7.3). For a variation of L, F_{ST} and N the χ^2-test and both GC methods show a similar power as before. However, the power of *P-STRAT* is now in most of the cases higher than the power of *P-CHIPOP* and increases if correct assignment of the individuals to the subpopulations is possible. The power of *P-CHIPOP* is in general comparable to the χ^2-test but slightly lower. Since the proportion of subpopulation 2 is larger with increasing RR, the power of the χ^2-test increases with higher RR. *P-STRAT* again is somewhat superior to *P-CHIPOP* for all different RR. For $K = 4$ a completely different result is obtained: here, even in the heterogeneous model, the median power of *P-CHIPOP* is 16.7% higher than the power of *P-STRAT*. Thus, also in a heterogeneous model the high degrees of freedom of *P-STRAT* yield to a power loss in comparison to *P-CHIPOP* and even to GC. Thus, although *P-CHIPOP* performs worse than in the homogeneous model it is the only SA statistic with a power always at least comparable but in most cases superior to GC.

7.3 Discussion

7.3.1 Summary

Many aspects are addressed in the simulations. First of all the simulations show that a correction for population stratification can be necessary in the realistic situation of large case-control studies with moderate population stratification because the variance inflation factor can be considerably large and the usual χ^2-test substantially inflated. However, for really small population structure as within Germany F_{ST}-values are even smaller than the simulated ones and the variance inflation often could be negligible depending on sample size and disease risks in the subpopulations.

The P-EM algorithm infers population structure quite well for the simulated data sets and the general population structure is already detected with 100 marker loci in most situations if $F_{ST} > 0.005$. However, for more subtle population structure including a very small F_{ST} and a larger number of subpopulations 100 marker loci definitely will not be sufficient as the simulations indicate. The standard EM algorithm shows the expected deviation of the estimated subpopulation proportions to the true proportions. Consequently the likelihood ratio test based on the standard EM algorithm is substantially inflated if the subpopulations cannot be correctly estimated. It is likely that the simulation results would not be very different applying a Bayesian clustering approach for discrete subpopulations instead of an EM algorithm since only discrete subpopulations were simulated. Thus, it can be concluded that the Structured Association approach as proposed by Pritchard et al. (2000a,b) is likely to show the same inflated type-I error rate.

A second question is which test statistic to use for Structured Association. For the likelihood ratio test as proposed by Pritchard et al. (2000b) it does not seem to be problematic to use the asymptotic χ^2-distribution since the test has the correct type-I error rate on average. But as expected, the likelihood ratio test statistic has a lack of power in situations where homogeneous allelic relative risks in the subpopulations are modelled. The power decreases with an increasing number of subpopulations. Surprisingly, also in the heterogeneous model the likelihood ratio test has a reduced power for a larger number of subpopulations. Since we are especially interested in moderate population structure a small variation of allelic relative risks across subpopulations is realistic but not a complete change. Thus, we recommend to apply the Wald test that we derived here instead of the likelihood ratio test. The main advantage of the Wald test is that allele frequency differences are averaged over subpopulations.

In comparison to Structured Association we also investigated Genomic Control theoretically and in simulations. We found that the median type-I error rate is in general close to the nominal level. However, Genomic Control shows a large variation dependent on the concrete marker set. We also theoretically calculated the variation of the observed type-I error rate for the mean-based test statistic and showed that this variation does not depend on the variance inflation factor itself, but only on the number of null loci. This is a weakness of Genomic Control in comparison to Structured Association where better clustering results are obtained with increasing F_{ST}. A second disadvantage of Genomic Control is the power loss in comparison to Structured Association if population structure and the differences in the disease prevalences between the subpopulations increase. Surprisingly, in our simulations Structured Association is also at least comparable to Genomic Control in situations with small population structure or only little genotyping of null loci regarding type-I error rate as well as power.

7.3.2 Comparison to other simulation studies

There are several other simulations published assessing the impact of population stratification and the possibilities to correct for it. Among these, there are a couple of simulations investigating the properties of Genomic Control (Bacanu et al., 2000; Marchini et al., 2004; Shmulewitz et al., 2004; Devlin et al., 2004). Bacanu et al. (2000) showed that in medium-sized case-control studies with less than 1000 individuals the variance inflation usually is not very large and thus the power of Genomic Control is high, especially in comparison to family-based association studies. Shmulewitz et al. (2004) investigated if the type-I error rate of Genomic Control is maintained including a small set of markers with large frequency differences into the set of null loci. In this case, the mean-based estimator for the variance inflation performs better than the median-based estimator. Marchini et al. (2004) investigated the type-I error rate of Genomic Control for large case control studies and moderate population stratification based on real marker data for a few number of individuals and an extrapolation to larger sample sizes. The main conclusion is that Genomic Control tests may be anticonservative for a small number of loci. Devlin et al. (2004) showed that this is only a consequence of applying the median-based estimator for the inflation factor instead of the mean-based estimator and the F-test (see section 4.2.1). In our simulations we confirm the main result from these recent simulations about Genomic Control that the mean-based estimator for the variance inflation factor should be preferred.

However, there are only few simulations investigating the differences between several test statistics correcting for stratification (Pritchard and Donnelly, 2001; Devlin et al., 2001a; Zhu et al., 2002; Chen et al., 2003). Pritchard and Donnelly (2001) compared Genomic Control and their own Structured Association method *STRAT* also simulating data in the beta-binomial model. These simulations are limited to a fixed parameter configuration corresponding to an inflation factor of $\lambda = 1.24$ and only the number of loci is varied. In the simulated scenario the average type-I error rate is close to the nominal level for both approaches and the power of Genomic Control and Structured Association is comparable. However, we show in our simulations that the results of Pritchard and Donnelly (2001) only cover a small number of possible simulation scenarios and cannot be automatically generalized. Furthermore we point out that the results are dependent on the concrete version of Structured Association. Devlin et al. (2001a) investigated the performance of the Structured Association test *STRAT* for a concrete simulation scenario consisting of many subpopulations. Although part of the population structure could not be detected, the observed type-I error rate was close to the nominal level. Their analysis suggests that a finite approximation to the structure present in realistic populations can protect at least against substantial confounding. However, a more systematic analysis is missing. Zhu et al. (2002) and Chen et al. (2003) simulated case-control data for discrete or admixed populations from different continents based on allele frequency data extracted

from a database. In their simulations *STRAT* also has an inflated type-I error rate in some cases which is likely to be explained by not including phenotype information in the clustering step. These simulations are quite unrealistic because cases and controls should be sampled from a population which is as homogeneous as possible. If at all different ethnicities are included these should be recorded and adjusted for in the analysis.
Thus, the main advantage of our simulation study is that it is on the one hand more general and on the other hand more realistic than others. The F_{ST}-values, sample sizes and the corresponding variance inflation factors are chosen to simulate realistic situations of large case-control studies with moderate population stratification (see chapter 6). Most of the other simulations concentrate either on extreme population stratification which is unlikely to appear in case-control studies (Zhu et al., 2002; Chen et al., 2003; Shmulewitz et al., 2004) or on small sample sizes which are not sufficient to detect weak associations (Bacanu et al., 2000; Pritchard and Donnelly, 2001; Zhu et al., 2002; Chen et al., 2003). A further strength of the study is that many parameter sets are investigated: the number of null loci, the fixation index F_{ST}, the sample size of the study, the disease prevalences and the number of subpopulations are varied. Only in the simulations of Marchini et al. (2004); Shmulewitz et al. (2004) investigating Genomic Control a comparable number of different parameters is studied. The other simulations are based on one or a few fixed parameter sets (Devlin et al., 2001a) and only the number of null loci is varied systematically (Bacanu et al., 2000; Pritchard and Donnelly, 2001; Chen et al., 2003). A third advantage of our simulation study is the large number of simulations for each parameter configuration which allows us to assess the variation of type-I error rate and power over the different multilocus marker sets. This variation is investigated systematically in no other simulation study up to date, the only aspect which is considered elsewhere is the variation of the inflation factor itself (Bacanu et al., 2000; Reich and Goldstein, 2001).

7.3.3 Limitations of the simulations

Our different simulation technique and increased complexity of our simulations lead to a more critical view of Genomic Control and a more positive view of Structured Association than obtained in other simulations (Bacanu et al., 2000; Shmulewitz et al., 2004; Devlin et al., 2004; Pritchard and Donnelly, 2001). There are also some weaknesses of our simulation study which may hide the problems of Structured Association. One of the disadvantages is that the multilocus marker data are simulated in the same model which is applied for inference on population structure via EM algorithm. We assume that it plays a minor role which simulation technique is used. The beta-binomial model should be appropriate for simulating diallelic multilocus marker data from discrete subpopulations. It is a commonly applied simulation technique, e.g. Pritchard and Donnelly (2001) simulate data in the same form as here. Marchini et al. (2004) use this simulation technique based on real data estimates for global allele frequencies and F_{ST}-values and show the good fit of

the model as well. However, an important weakness of our simulations is that admixture of populations is not simulated although it is the realistic situation for almost all populations worldwide. If an admixed population is clustered into discrete subpopulations a residual error remains whatever number of loci is used for clustering the individuals. This problem is not addressed in our simulations. Furthermore we did not consider a high number of subpopulations as well as really small population structure as for example existing within Germany. These points need further investigations in the future. However, more complicated population structure than simulated here is probably more problematic for Structured Association whereas Genomic Control is independent of the sort of population structure. Nevertheless we come to the conclusion that at least for simple population structure Structured Association if applied correctly is superior to Genomic Control.

8 Summary and outlook

Theoretical considerations as well as simulations showed that the problem of population stratification in case-control studies is a very complicated topic where many different aspects have to be considered. Thus, at the end of the thesis we want to give an outlook on research questions still to be investigated especially regarding the two main approaches Genomic Control and Structured Association.

We mainly focussed on proposing a new Structured Association approach and discussing further Structured Association approaches (see chapter 5). Section 5.4 contains a detailed discussion of our main theoretical results. We showed that Structured Association has to be applied with a clustering algorithm conditioning on the phenotype if subsequently a test statistic based on the likelihood function for the genotype data at the candidate locus is applied. Otherwise a systematic bias is introduced when estimating the subpopulation proportions within cases and controls. As an appropriate clustering algorithm we proposed the phenotype-dependent EM algorithm. Thus, in our theoretical development we concentrated on the idealistic situation that the total population is assumed to consist of discrete subpopulations. There are approaches which incorporate admixture (Pritchard et al., 2000a; Hoggart et al., 2003; Shmulewitz et al., 2004) but this is still a new field of research which we expect to develop further in the future.

As a second step of Structured Association we developed a Wald test which is theoretically designed for the situation that population stratification only acts as a confounder but not as an effect modifier. It can be applied for testing a diallelic candidate marker based on the inferred structure. This is a very common situation, but often further research questions shall be investigated. In many studies several SNPs being in LD with each other are genotyped on the same gene and a haplotype-based test statistic would be appropriate testing all SNPs simultaneously. Moreover, we focussed on statistical hypothesis testing but do not consider effect estimation in the presence of hidden population stratification. It is still an open task to derive estimators for subpopulation-specific odds ratios as well as for a common allelic odds ratio in the subpopulations based on the posterior subpopulation probabilities or admixture proportions. In addition, formulae for the corresponding confidence intervals have to be developed.

Our association test follows the classical approach which is based on the likelihood for the genotype data given the phenotype data. As already discussed in section 5.2.4 a logistic regression model could also be applied to test for association following a different likelihood approach. A logistic regression model has the advantage that an effect estimate for a common odds ratio automatically is obtained and that it is easy to incorporate further covariates in the model. Details of this model could be worked out in the future.

Our association test is developed for a candidate gene approach where additional markers are genotyped to control for population stratification. However, in the last years whole

genome association scans became more and more popular. SNP chips have been developed where about 500,000 SNPs of one individual can be genotyped simultaneously. New approaches have to be developed to handle such large amounts of data. In order to infer population structure from such a large number of SNPs current models have to be extended to allow for linkage disequilibrium between adjacent markers. In whole genome scans the question is to simultaneously infer population structure and find the SNPs which show a significant association to the phenotype of interest. Only the approach of Hoggart et al. (2003) is designed to investigate this kind of research question but the performance is not well tested yet.

Besides a theoretical comparison of the methods it also is important to analyze appropriate data sets with the different association tests. However, as discussed in chapter 6 there are only few at least medium-sized case-control studies where additional marker have been genotyped. Of course, such data sets are not freely available. The analysis of the German Genomic Control Study revealed that within Germany there is too little population structure to apply Structured Association successfully. However, the variance inflation can be considerable if cases and controls are sampled from different German regions (see section 6.3).

We also investigated the performance of different Structured Association tests in simulations and compared it to Genomic Control (see chapter 7). The results of the simulations are discussed in detail in section 7.3. Regarding the comparison of the Structured Association test statistics it turned out that the Wald test statistic had a substantially higher power than the likelihood ratio test statistic for a larger number of subpopulations. This also held true for simulating an effect modification with different allelic relative risks in the subpopulations but the same high risk allele. Thus, to adjust for confounding by population stratification we propose the Wald test statistic. We covered several realistic simulation scenarios with many replications to come to a general conclusion about the performance of Structured Association and Genomic Control for large case-control studies with small to moderate population stratification. At least for simple population structure as simulated here Structured Association is superior to Genomic Control. A disadvantage of Genomic Control turned out to be the large variation in estimating the variance inflation factor as well as the power loss if population structure increased. However, as already discussed in section 7.3.3 there are still several aspects which could be investigated in further simulations. Among these the most important is to incorporate admixture in the simulation model. More complicated population structure than simulated here could turn out to be more problematic for Structured Association than for Genomic Control which is independent of the type of population structure. Thus, other simulations have to be carried out before coming to a final conclusion.

A Appendix

A.1 Notation

In this section the notation which is used throughout the thesis is summarized. In general bold letters are used for vectors. Random variables are usually denoted with capital letters and their realizations with small letters. The following general notation is used:

$\boldsymbol{X}$	multilocus genotype marker data which determine population structure
$\boldsymbol{\varphi}$	allele frequencies for the multilocus genotype marker data
$\boldsymbol{G}$	genotype data at the candidate locus
$\boldsymbol{p}$	allele frequencies at the candidate locus
$\boldsymbol{\pi}$	subpopulation proportions

The detailed list of symbols is following here in the same order as it is introduced in the text. In the first column the scalars are given and in the following column the corresponding vectors if defined at the same place or somewhere later in the text. The notation introduced in one chapter is used throughout the whole thesis unless otherwise mentioned.

Chapter 2:

X_{ilj}	$\boldsymbol{X}_i, \boldsymbol{X}$	allele from individual i, at DNA-strand j, at locus l, coded as 0=B,1=b
X_{il}		genotype from individual i, at locus l, coded as 0,1,2
φ_l		global allele frequency at locus l
$D_{ll'}$		linkage disequilibrium (LD) between the loci l and l'
$\Delta_{ll'}$		standardized measure for LD between the loci l and l'
$\theta_{ll'}$		recombination fraction between the loci l and l'
R_l		number of possible alleles at locus l for multiple alleles coded as $B_1, \dots B_{R_l}$
$X_{ilj}^{(r)}$	$\boldsymbol{X}_{ilj}$	indicator variable if allele r is present in individual i, at strand j, at locus l
$X_{il}^{(r)}$	$\boldsymbol{X}_{il}$	count of allele r in individual i at locus l
$\varphi_l^{(r)}$	$\boldsymbol{\varphi}_l$	global frequency for allele r at locus l
F_i		Wright's coefficient of inbreeding for individual i
$f_{ii'}$		kinship coefficient between individuals i and i'
K		number of subpopulations, referred to as $S_1, \dots S_K$
F_{IT}		global inbreeding coefficient
F_{ST}		fixation index
F_{IS}		local inbreeding coefficient
φ_{kl}	$\boldsymbol{\varphi}_k, \boldsymbol{\varphi}$	allele frequency in subpopulation k at locus l
π_k	$\boldsymbol{\pi}$	proportion of subpopulation k in the total population
Z_i	$\boldsymbol{Z}$	subpopulation of individual i
F_k		distance of subpopulation k to the ancestral population
N_k		number of individuals in sample from subpopulation k (S_k)
$\varphi_{kl}^{(r)}$	$\boldsymbol{\varphi}_{kl}$	frequency for allele r in S_k at locus l
q_{ik}^A	$\boldsymbol{q}_i^A$	proportion of individual i's genome originated in S_k in the admixture model
N		number of individuals in a population sample
L		number of diallelic marker loci

Chapter 3:

G		genotype of any individual coded as 0, 1 or 2
a, u		possible disease status, a for cases (affecteds), u for controls (unaffecteds)
Y		phenotype of any individual coded as a or u
$f_g^{(y)}$	$\boldsymbol{f}^{(y)}$	probability for genotype g given disease status y
ϕ_g^G		penetrance for genotype g
ψ_g^G		genotypic relative risk for genotype g in comparison to genotype 0
ξ_g^G		genotypic odds ratio for genotype g in comparison to genotype 0
f_g^T	$\boldsymbol{f}^T$	probability for genotype g in the total population
ϕ		disease prevalence
p^T		allele frequency in the total population
$p^{(y)}$		allele frequency given disease status y
ξ		allelic odds ratio
ψ		allelic relative risk
N		total number of individuals, refers to the case-control sample here
$N^{(y)}$		number of individuals with disease status y
$r_g^{(y)}$		number of individuals with genotype g and disease status y
r_g		number of individuals with genotype g
$s^{(y)}$		number of B-alleles in individuals with disease status y
s		number of B-alleles in the total sample
G_i	$\boldsymbol{G}$	genotype of individual i coded as $0, 1, 2$
G_{ij}		allele of individual i at DNA-strand j coded as 0, 1
Y_i	$\boldsymbol{Y}$	phenotype of individual i
p		overall allele frequency under the null hypothesis of no association
$\pi_k^{(y)}$	$\boldsymbol{\pi}^{(y)}$	proportion of the subpopulation k within individuals of phenotype y
$N_k^{(y)}$		number of individuals with phenotype y from subpopulation k in the sample
$p_k^{(y)}$	$\boldsymbol{p}^{(y)}$	allele frequency in subpopulation k for phenotype y
p_k	$\boldsymbol{p}$	allele frequency in subpopulation k under the null hypothesis
N_k		number of individuals in subpopulation k
$s_k^{(y)}$		number of B-alleles in subpopulation k for phenotype y
s_k		number of B-alleles in subpopulation k
c_k		weight for subpopulation k in the Cochran-Mantel-Haenszel test
ξ_k	$\boldsymbol{\xi}$	allelic odds ratios in subpopulations k
$\xi^{(0)}$		common odds ratio in all subpopulations
v_{iq}	$\boldsymbol{v}_i$	vector of exposures in the logistic regression model
ϕ_i		probability of individual i being a case in the case control sample
α		intercept of the regression
β_q	$\boldsymbol{\beta}$	regression coefficient for exposure q
ξ_q		odds ratio for a binary exposure q
g_{ig}^G	$\boldsymbol{g}_i^G$	dummy coding for the genotype of individual i
β_g^G	$\boldsymbol{\beta}^G$	regression coefficient for the genotype g (dummy coding)
β		regression coefficient for the genotype
η_k	$\boldsymbol{\eta}$	regression coefficient for subpopulation k in the logistic regression
z_{ik}^D	$\boldsymbol{z}_i^D$	dummy coding for the subpopulation of individual i
η_k^D	$\boldsymbol{\eta}^D$	regression coefficient for subpopulation k (dummy coding)

Chapter 4:

λ		variance inflation factor
L		number of marker loci, refers to additionally genotyped marker loci here
T^2		test statistic Genomic Control is based on
T_l^2		same test statistic for marker locus l

Chapter 5:

q_{ik}^*		posterior probability of individual i of being from S_k given $\boldsymbol{x}_i$
q_{ik}		posterior probability of individual i of being from S_k given $\boldsymbol{x}_i$, y_i
$\gamma_k^{(y)}$		risk for phenotype y in S_k in the study sample
ν_{il}		variable which indicates if genotypes are available at locus l for individual i
D		number of derived classes in the discrete admixture model
ζ_{dk}	$\boldsymbol{\zeta}_d$	proportion of the genome from the ancestral S_k for the derived class d
Z_i^{DC}		derived class of an individual i
Z_{ilj}^A		ancestral subpopulation of allele X_{ilj}
π_k		proportion of alleles from S_k in the sample in the admixture model
α_0		admixture parameter for the amount of admixture in the sample
$\pi_k^{(y)}$		proportion of alleles from S_k given phenotype y in the admixture model
$\alpha_0^{(y)}$		amount of admixture in individuals of phenotype y
q_{ik}	$\boldsymbol{q}_i, \boldsymbol{q}$	posterior probability or admixture proportion given y_i in the association test
	$\boldsymbol{p}^1$	allele frequencies at the candidate locus under the alternative
	$\boldsymbol{p}$	allele frequencies at the candidate locus under H_0 (see chapter 3)
Z_{ij}^C	$\boldsymbol{Z}^C$	ancestral subpopulation of allele G_{ij} at the candidate locus
	$\boldsymbol{I}(\boldsymbol{p}^1)$	information matrix
	$\boldsymbol{c}$	contrast vector for Wald-type statistic
q_{ik}	$\boldsymbol{q}_i, \boldsymbol{q}$	posterior prob. of S_k given $\boldsymbol{x}_i$ in the logistic regression

Chapter 7:

ψ_k	allelic relative risk in S_k
RR	relative risk for disease in a further subpopulation compared to subpop. 1

A.2 Statistical Theory

A.2.1 Likelihood based tests of significance

This section gives a summary of the likelihood based test theory applied for standard tests of association and also for the derivation of the new Wald test correcting for population stratification. The description of the theory is mainly based on Serfling (1980) and Kendall and Stuart (1973). Let $X_1, \ldots, X_N$ be iid random variables with a density or probability mass function $f_{\boldsymbol{\theta}}(x)$ belonging to a family $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ where $\boldsymbol{\Theta}$ is an open subset in $\mathbb{R}^q$. The *likelihood function* then gives the total likelihood of the sample $x_1, \ldots, x_N$ under the assumed model

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f_{\boldsymbol{\theta}}(x_i).$$

The *maximum likelihood estimate* of $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}_N$, is the vector which maximizes the likelihood. This vector is most easily determined using the log-likelihood function $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. If the log-likelihood function is a twice differentiable function, a necessary condition for a local turning point is that the estimate is a solution of the maximum likelihood estimating equations, i.e.

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, \ldots, q.$$

In many cases the solution of these estimating equations requires an iterative algorithm as for example the EM algorithm described in appendix A.2.2. A sufficient condition for the local turning point to be a maximum is that the matrix

$$\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right)_{j,k=1,\ldots q}$$

is negative definite.
An important quantity with respect to variance estimation is *Fisher's Information matrix* which is defined as

$$I_N(\boldsymbol{\theta}) = \mathrm{E} \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_k} \right)_{j,k=1,\ldots q}$$

and can also be written as

$$I_N(\boldsymbol{\theta}) = -\mathrm{E} \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right)_{j,k=1,\ldots q}.$$

The information matrix quantifies the expected amount of information in the sample concerning the true vector $\boldsymbol{\theta}$ since the second derivatives describe the curvature of the log-likelihood in the neighbourhood of $\boldsymbol{\theta}$.
Under certain regularity conditions the asymptotic distribution of the maximum likelihood estimate can be derived.

Proposition A.1. *Under certain regularity conditions as described in Kendall and Stuart (1973); Lehmann (1983) a maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_N$ exists with a probability tending to 1 for $N \to \infty$ so that*

1. *$\widehat{\boldsymbol{\theta}}_N$ is a consistent estimator for $\boldsymbol{\theta}$.*
2. *$\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})$ is asymptotically $N(\mathbf{0}, N I_N(\boldsymbol{\theta})^{-1})$ distributed.*

It follows from the second statement that the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ is *asymptotically efficient*, i.e. for any vector $\boldsymbol{c} \in \mathbb{R}^q$ the asymptotic variance of $\sqrt{N}\boldsymbol{c}'(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})$ is minimal. This result is derived from the Cramér-Rao lower bound for the variance of an unbiased estimate $\widehat{\boldsymbol{\phi}}$ for the vector $\boldsymbol{\phi} \in \mathbb{R}^q$. For a any linear combination $\boldsymbol{c}'\widehat{\boldsymbol{\phi}}$ the Cramér-Rao lower bound for the variance is given by

$$\mathrm{Var}\,(\boldsymbol{c}'\widehat{\boldsymbol{\phi}}) \geq \boldsymbol{c}'\boldsymbol{I}(\boldsymbol{\phi})^{-1}\boldsymbol{c}.$$

The maximum likelihood theory can be used to construct three different types of large sample tests, the Wald test, the likelihood ratio test and the score test. The first two test statistics are applied here and thus explained subsequently. A null hypothesis H_0 to be tested is specified as a subset $\boldsymbol{\Theta}_0$ of $\boldsymbol{\Theta}$ where $\boldsymbol{\Theta}_0$ is determined by a set of $r \leq q$ restrictions given by equations

$$R_i(\boldsymbol{\theta}) = 0, \quad 1 \leq i \leq r.$$

In the case of a *composite* hypothesis the set $\boldsymbol{\Theta}_0$ contains more than one element and then necessarily $r < q$. A special case of a composite null hypothesis is a linear hypothesis of the form $H_0 : \boldsymbol{C}'\boldsymbol{\theta} = \boldsymbol{d}$ where $\boldsymbol{C}'$ is a $r \times q$ contrast matrix of rank r and $\boldsymbol{d}$ is the $r \times 1$ solution vector. To calculate the *Wald test* (Wald, 1943) for a linear hypothesis, the large sample variance $NI_N(\boldsymbol{\theta})^{-1}$ has to be estimated by a consistent estimator $N\widehat{\boldsymbol{\Sigma}}_N$.

Proposition A.2 (Wald test)**.** *Under the same regularity conditions as before the Wald-type test statistic for a linear hypothesis*

$$(\boldsymbol{C}'\widehat{\boldsymbol{\theta}}_N - \boldsymbol{d})'[\boldsymbol{C}'\widehat{\boldsymbol{\Sigma}}_N\boldsymbol{C}]^{-1}(\boldsymbol{C}'\widehat{\boldsymbol{\theta}}_N - \boldsymbol{d})$$

is asymptotically χ^2_r-distributed.

A consistent estimator $N\widehat{\boldsymbol{\Sigma}}_N$ is given by $NI_N(\widehat{\boldsymbol{\theta}})^{-1}$ which estimates the variance under the alternative. However, tests which estimate the variance under the null hypothesis are more efficient although all such tests are asymptotically equivalent. Thus, if possible the variance should be estimated under the null hypothesis.
For the *likelihood ratio test* variance estimation is not necessary. The asymptotic distribution of the likelihood ratio test statistic was originally derived by Wilks (1938).

Proposition A.3 (Likelihood ratio test)**.** *Let the likelihood ratio test statistic for a general composite hypothesis be given as*

$$\Lambda_N = \frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})}.$$

Under the same regularity conditions as before $-2\log(\Lambda_N)$ is asymptotically χ^2_r-distributed.

A.2.2 The EM algorithm

The EM algorithm (Dempster et al., 1977) is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution if the data are incomplete or have missing values. The EM algorithm can be applied if optimizing the likelihood is analytically impossible but if the likelihood function can be simplified by assuming the existence of additional but missing or hidden parameters.
Let $\boldsymbol{X} = (\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_N)'$ be a vector of iid random vectors $\boldsymbol{X}_i$ which have a density

or probability mass function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ with the unknown parameter vector $\boldsymbol{\theta}$. The original likelihood function based on the observed incomplete data $\boldsymbol{x}$

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{N} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

is referred to as *incomplete data likelihood*. We assume the existence of a missing data vector $\boldsymbol{z}$ and a joint density or probability mass function

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) = f_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x}).$$

The *complete data likelihood* can be defined as $L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{Z}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Z})$ which is in fact a random variable since the missing information $\boldsymbol{Z}$ is unknown.
The EM algorithm (expectation maximization algorithm) is defined based on the complete data likelihood.

EM algorithm: The following two steps have to be repeated iteratively for $t = 1, 2 \ldots$:
E-step: The expected value of the complete data log-likelihood $\log L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{Z})$ with respect to the missing data $\boldsymbol{Z}$ given the observed data $\boldsymbol{x}$ and the current parameter estimates $\boldsymbol{\theta}^{(t)}$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}\left[\log L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{Z})|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}\right]$$

has to be determined. To evaluate the expectation the density or probability mass function of the missing data $f(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})$ has to be known.
M-step: The second step is to choose $\boldsymbol{\theta}^{(t+1)}$ as the vector which maximizes the expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect $\boldsymbol{\theta}$.

Proposition A.4. *Each iteration of the EM algorithm monotonically increases the incomplete data log-likelihood and the EM algorithm is guaranteed to converge to a local maximum of the likelihood function.*

Proof: see Dempster et al. (1977). □

References

Abecasis, G. R., Noguchi, E., Heinzmann, A., Traherne, J. A., Bhattacharyya, S., Leaves, N. I., Anderson, G. G., Zhang, Y., Lench, N. J., Carey, A., Cardon, L. R., Moffatt, M. F., and Cookson, W. O. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet*, 68(1):191–197.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

Bacanu, S. A., Devlin, B., and Roeder, K. (2000). The power of genomic control. *Am J Hum Genet*, 66(6):1933–1944.

Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12.

Bamshad, M. J., Wooding, S., Watkins, W. S., Ostler, C. T., Batzer, M. A., and Jorde, L. B. (2003). Human population genetic structure and inference of group membership. *Am J Hum Genet*, 72(3):578–589.

Böhning, D. (2000). *Computer-Assisted Analysis of Mixtures and Applications. Meta-analysis, disease mapping and others*. Chapman & Hall, Boca Raton.

Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, Berkeley CA.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G., and Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nat Genet*, 37(8):868–872.

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1996). *The History and Geography of Human Genes*. Princeton University Press, Princeton, New Jersey.

Chen, H.-S., Zhu, X., Zhao, H., and Zhang, S. (2003). Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet*, 67(Pt 3):250–264.

Clayton, D. (2000). Population association. In Balding, D. J., Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*. Wiley.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics*, 10:417–451.

Collett, D. (2003). *Modelling binary data*. Chapman & Hall, Boca Raton.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

Devlin, B., Bacanu, S.-A., and Roeder, K. (2004). Genomic Control to the extreme. *Nat Genet*, 36(11):1129–1130. Comment.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

Devlin, B., Roeder, K., and Bacanu, S. A. (2001a). Unbiased methods for population-based association studies. *Genet Epidemiol*, 21(4):273–284.

Devlin, B., Roeder, K., and Wasserman, L. (2001b). Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol*, 60(3):155–166.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Ewens, W. J. and Spielman, R. S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet*, 57(2):455–464.

Excoffier, L. (2000). Analysis of population subdivision. In Balding, D. J., Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*. Wiley.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nat Genet*, 36(4):388–393.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Gillespie, J. H. (1998). *Population Genetics - A concise guide*. The Johns Hoppkins University Press, Baltimore, Maryland.

Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J., and Stefansson, K. (2005). An Icelandic example of the impact of population structure on association studies. *Nat Genet*, 37(1):90–95.

Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G., and McKeigue, P. M. (2003). Control of confounding of genetic associations in stratified populations. *Am J Hum Genet*, 72(6):1492–1504.

Human Genome Management Information System (2003). *Genomics and its Impact on Science and Society: The Human Genome Project and Beyond*. Oak Ridge National Laboratory, Oak Ridge, Tennesse.

Karr, A. F. (1993). *Probability*. Springer, New York.

Kendall, M. G. and Stuart, A. (1973). *Advanced Theory of Statistics*, volume Volume 2, Inference and Relationship. Griffin, London.

Köhler, K. and Bickeböller, H. (2006). Case-Control Association Tests Correcting for Population Stratification. *Ann Hum Genet*, 0:–. In print, doi:10.1111/j.1529-8817.2005.00214.x.

Knapp, M. (1999). The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/ disequilibrium test. *Am J Hum Genet*, 64(3):861–870.

Knowler, W. C., Williams, R. C., Pettitt, D. J., and Steinberg, A. G. (1988). Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet*, 43(4):520–526.

Lachin, J. M. (2000). *Biostatistical Methods*. Wiley, New York.

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265(5181):2037–2048.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

Li, C. C. (1969). Population subdivision with respect to multiple alleles. *Ann Hum Genet*, 33(1):23–29.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, 22:719–748.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nat Genet*, 36(5):512–517.

Maynard Smith, J. (1989). *Evolutionary Genetics*. Oxford University Press, Oxford.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

McPherson, J. D. et al. (2001). A physical map of the human genome. *Nature*, 409(6822):934–941. The International Human Genome Mapping Consortium.

Miettienen, O. S. (1985). *Theoretical Epidemiology*. Wiley, New York.

Nagylaki, T. (1998). Fixation indices in subdivided populations. *Genetics*, 148(3):1325–1332.

Nei, M. (1977). F-statistics and analysis of gene diversity in subdivided populations. *Ann Hum Genet*, 41(2):225–233.

Nei, M. and Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. *Ann Hum Genet*, 47 (Pt 3):253–259.

Noether, G. E. (1955). On a theorem of Pitman. *Ann. Math. Statist.*, 26:64–68.

Paetkau, D., Waits, L. P., Clarkson, P. L., Craighead, L., and Strobeck, C. (1997). An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. *Genetics*, 147(4):1943–1957.

Parkin, D. M., Whelan, S. L., Ferlay, J., Teppo, L., and Thomas, D. B., editors (2002). *Cancer Incidence in Five Continents*, volume VIII of *IARC Scientific Publications*. IARC Press, Lyon.

Pritchard, J. K. and Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. *Theor Popul Biol*, 60(3):227–237.

Pritchard, J. K. and Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, 65(1):220–228.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b). Association mapping in structured populations. *Am J Hum Genet*, 67(1):170–181.

Purcell, S. and Sham, P. (2004). Properties of structured association approaches to detecting population stratification. *Hum Hered*, 58(2):93–107.

Rabinowitz, D. and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered*, 50(4):211–223.

Radhakrishna, G. E. (1965). Combination of results from several 2 x 2 contingency tables. *Biometrics*, 21:86–98.

Reich, D. E. and Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol*, 20(1):4–16.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.

Risch, N. and Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res*, 8(12):1273–1288.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.

Rothman, K. J. (1986). *Modern Epidemiology*. Little, Brown and Company, Boston.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261.

Satten, G. A., Flanders, W. D., and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet*, 68(2):466–477.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Sham, P. (1998). *Statistics in Human Genetics*. Arnold, London.

Shmulewitz, D., Zhang, J., and Greenberg, D. A. (2004). Case-control association studies in mixed populations: correcting using genomic control. *Hum Hered*, 58(3-4):145–153.

Spielman, R. S. and Ewens, W. J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet*, 62(2):450–458.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 52(3):506–516.

Steffens, M., Lamina, C., Illig, T., Bettecken, T., Vogler, R., Entz, P., Suk, E.-K., Toliat, M. R., Klopp, N., Caliebe, A., König, I. R., Köhler, K., Lüdemann, J., Lacava, A. D., Fimmers, R., Ziegler, A., Wolf, A., Krawczak, M., Nürnberg, P., Hampe, J., Schreiber, S., Meitinger, T., Wichmann, H.-E., Roeder, K., Wienker, T. F., and Baur, M. P. (2006). SNP-based analysis of genetic substructure in the german population. Submitted.

Stein, L. D. (2004). Human genome: end of the beginning. *Nature*, 431(7011):915–916. Comment.

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S. L. R., Zhu, X., Brown, A., Pankow, J. S., Province, M. A., Hunt, S. C., Boerwinkle, E., Schork, N. J., and Risch, N. J. (2005). Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet*, 76(2):268–275.

Thomas, D. C. and Witte, J. S. (2002). Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev*, 11(6):505–512. Editorial.

Titterington, D. M., Smith, A. F. M., and Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Wacholder, S., Rothman, N., and Caporaso, N. (2002). Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev*, 11(6):513–520. Comment.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.*, 54:426–482.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38:1358–1370.

Weir, B. S. and Hill, W. G. (2002). Estimating F-statistics. *Annu Rev Genet*, 36:721–750.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9:60–62.

Woolf, B. (1955). On estimating the relationship between blood group and disease. *Ann. Human. Genet.*, 19:251–253.

Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist*, 56:330–338.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15:323–354.

Zhu, X., Zhang, S., Zhao, H., and Cooper, R. S. (2002). Association mapping, using a mixture model for complex traits. *Genet Epidemiol*, 23(2):181–196.

Index

Curriculum Vitae

Geburtstag	13.05.1976
Geburtsort	Göttingen
1982 - 1986	Grundschule Diemarden
1986 - 1988	Orientierungsstufe Jahnschule in Göttingen
1988 - 1995	Theodor-Heuss-Gymnasium in Göttingen
Okt. 95 - Nov. 01	Mathematikstudium, Studienrichtung Wirtschaftsmathematik an der Georg-August-Universität Göttingen Diplomarbeit: Unverzerrte Schätzer für die relativen Effekte im nichtparametrischen gemischten Modell Betreuer der Diplomarbeit: Prof. Brunner Abschluss: Diplom
Feb. 00 - Jan. 01	Nebentätigkeit als studentische Hilfskraft in der Abteilung Medizinische Statistik der Universität Göttingen
seit Dez. 2001	wissenschaftliche Mitarbeiterin in der Abteilung Genetische Epidemiologie der Universität Göttingen bei Frau Prof. Bickeböller mit dem Ziel der Promotion
seit Okt. 2002	Teilnahme am interdisziplinären Promotionsstudiengang "Angewandte Statistik und empirische Methoden" der Universität Göttingen