English Summary

This dissertation describes new statistical methods designed to improve the power of genetic association studies. Of particular interest are studies with a response-selective sampling design, i.e. case-control studies of unrelated individuals and case-control studies of family members. In the pages that follow, we detail novel statistical methods that (a) take advantage of information available in the distribution of the covariates in case-control studies by modeling the ascertainment process; (b) incorporate information from both family-based studies and case-control studies of unrelated individuals; (c) use "richer" models of the relationship between genetic variants and phenotypes, compared to models used in standard genetic association studies; and (d) integrate different types of data, such as genomic, epigenomic, transcriptomic and environmental information. Together, these methods will improve the ability of the genetics community to identify the genetic basis of complex human phenotypes.

Chapter 1 provides a general introduction to existing methods for the statistical analysis of genetic association studies with response-selective sampling designs. We start by introducing the relevant terminology and the key concepts of genetic association studies. Next, we present and compare the two most popular response-selective sampling designs in genetic studies: case-control studies of unrelated individuals and case-control studies of family members. We proceed to explain the two main advantages of accounting for ascertainment in such studies: the potential increase in power to detect associations and proper secondary phenotype analysis.

The rest of the introduction is split in two parts. In the first part, we present three different likelihood approaches for modelling the ascertainment in family-based case-control studies. The first approach is based on the prospective likelihood, which models the distribution of phenotypes conditional on covariates and ascertainment. The second approach is the ascertainment-corrected joint likelihood, which models the joint distribution of phenotypes and covariates conditional on ascertainment. The last approach is the retrospective likelihood, which models the distribution of covariates conditional on phenotypes. The latter is also appropriate for the analysis of case-control data of unrelated individuals. The likelihoods are compared in terms of efficiency of parameter estimates and computational efficiency.

In the second part of the introduction we describe different models for the relation between the genetic variants and the phenotype. The current standard analysis protocol for genome wide association studies is to individually evaluate the relationship between each SNP and disease. However, most common complex diseases do not arise from a single genetic cause, but rather a combination of multiple genetic and environmental factors (i.e., they are polygenic). Here, we present alternative approaches, which more closely model the underlying biological mechanisms, such as jointly modeling multiple genetic variants, or jointly modeling genetic variants and intermediate cellular phenotypes.

Chapter 2 describes a novel method to improve the power of genetic association studies by combining data from multi-case family studies and twin studies and modeling the ascertainment process of such studies. In order to maximize efficiency in parameter estimation, inference about the parameters of interest is based on an ascertainment-corrected joint like-

lihood. To take into account the correlation of disease risks among family members, due to shared but unmeasured genetic or environmental factors, a family-specific random term is used.

Simulations and real data analysis show that this ascertainment-corrected joint likelihood combining family and twin data is more efficient for estimating the parameters of interest, as compared to a families-only ascertainment-corrected joint likelihood approach or a prospective likelihood approach which ignores the ascertainment. The combined approach, not only enhances the statistical power to detect direct offspring allelic effects, but also effects depending on maternal-offspring genotype combinations, such as non-inherited maternal antigen effects. The efficiency improvement of the joint likelihood over the prospective likelihood is higher when information is limited, i.e. when the families are small (three offspring per family) and ascertained such that at least two out of three offspring are affected. The efficiency improvement of the combined families-and-twins approach against the families-only approach is noticeably high when the sample size is small, i.e. the number of families in the study is 100 or less.

Chapter 3 considers an alternative haplotype based strategy to the current gold standard of marginal testing. Marginal tests based on individual SNPs have dominated association analyses in the past decade. Although single SNP analyses have led to the identification of hundreds of genetic variants associated with many complex diseases, greater power might be gained by using haplotype-based approaches to analyze multiple markers simultaneously. Haplotype-based association methods incorporate linkage disequilibrium (LD) information from multiple markers and can be more powerful for gene mapping than methods based on single SNPs. A limitation of haplotype-based methods is that the number of parameters increases exponentially with the number of SNPs, inducing a commensurate increase in the degrees of freedom and weakening the power to detect associations.

Here we consider a hierarchical linkage disequilibrium model for trait mapping that enables flexible testing strategies over a range of hypotheses: from single SNP analyses through the haplotype distribution tests. Many such models reduce d.f. and increase the power to detect associations. These models are based on a re-parametrization of the multinomial haplotype distribution, where every parameter corresponds to the joint cumulant of each possible subset of a set of loci. Extensive simulations and a real data analysis show that such tests, which make plausible restrictions on the parameter space, have often increased power against the unrestricted global haplotype test for association or the single-SNP tests.

Genetic studies aim to assess the association between genetic variants and common complex traits. For the analysis of such traits, two different methods can be used: linkage mapping and association mapping. In Chapter 4, we consider the trade-offs between these two methods. Linkage mapping methods are more powerful for identifying rare variants with large effect on disease susceptibility while association-mapping methods are more suitable for identifying more common variants with moderate effect sizes. However, SNPs typically have small effect sizes (common variants) or minor allele frequencies that are too small to reliably fit models (rare variants). If the rare variant effects were large, and the disease was not heterogeneous, they would have been found through previous family-based linkage studies. Thus, there may be a middle ground in which multiple rare variants of moderate to low effect size play a key role in the etiology of some diseases. Such situations might be ideal for combining linkage- and association-mapping.

We develop a two-part analysis in order to investigate the contribution that linkage-based methods, such as IBD mapping, can make to association mapping to identify rare variants in next-generation sequencing data. In the first part we use identity-by-descent (IBD) mapping to identify regions in which cases share more segments of IBD around a putative causal variant than do controls. In the second part we perform association-mapping

by using a two-stage mixed-effects model approach to summarize the SNP data within the regions identified in the first part and including them as covariates in the model for the phenotype. To increase our power to identify rare variants, we also include the number of rare variants per region as a covariate in the model. The method was applied to next-generation sequencing longitudinal family data from Genetic Association Workshop 18 and a significant association was identified.

Chapter 5 examines integrative omics, the joint analysis of outcome and multiple types of omics data, such as genomics, epigenomics and transcriptomics data. Integrative omics has emerged as a promising approach for powerful and biologically relevant association studies. These studies often employ a case-control design, and often include non-omics covariates, such as age and gender, that may modify the underlying omics risk factors. An open question is how to best integrate multiple omics and non-omics information to maximize statistical power in case-control studies that ascertain individuals based on the phenotype. Recent works on integrative omics have used prospective approaches, modeling case-control status conditional on omics and non-omics risk factors. Compared to univariate approaches, jointly analyzing multiple risk factors with a prospective approach increases power in non-ascertained cohorts. However, in case-control studies this is no longer the case and these prospective approaches often lose power compared to univariate approaches.

We present a novel statistical method for integrating multiple omics and non-omics factors that addresses these issues of power loss in case-control association studies. This method is based on a retrospective likelihood function that models the joint distribution of omics and non-omics factors conditional on case-control status. In order to model the distribution of the risk factors as efficiently as possible, knowledge about the correlation structure between risk factors in the population is exploited and parametric assumptions about the distribution of the risk factors are made. The new method provides accurate control of Type I error rate and has increased efficiency over prospective approaches in both simulated and real data. Efficiency gain is a function of the number of parameters used to model the distribution of the risk factors and the effect sizes of risk factors, with increased efficiency gain for continuous factors and for risk factors with large effect sizes.

Chapter 6 considers the problem of phenotype description. Sometimes an outcome is based on rating of multiple underlying features and might thereby be prone to inter-rater variability. In these cases the outcome definition can be made objective by learning a predictor for the outcome based on the underlying multivariate data. Potentially this can improve power of ensuing studies and improve the understanding of the outcome variable.

Here, we consider genetic syndromes as such a phenotype and 2D graph-data derived from facial images as features. We present a method for automated syndrome classification and visualization of the classifier. In order to optimize the classifier, we investigate a set of data transformations prior to analysis and their effect on classification accuracy in a high-dimensional setting. These transformations are low-variance in the sense that each involves only a fixed small number of input features. It is shown that classification accuracy can be improved when penalized regression techniques are employed, as compared to a principal component analysis pre-processing step.

A second goal is to visualize the resulting classifiers. We develop importance plots highlighting the influence of coordinates in the original 2D space. Features used for classification are mapped to coordinates in the original images and combined into an importance measure for each pixel. These plots assist in assessing plausibility of classifiers, interpretation of classifiers, and determination of the relative importance of different features.