Characterizing Association Parameters in Genetic Family-based Association Studies

Dissertation by

Stefan Böhringer

Institut für Humangenetik, Universität Duisburg-Essen,

Hufelandstr. 55, 45122 Essen, Germany

correspondence@s-boehringer.org

Submitted to the Department of Statistics

of the University of Dortmund

in Fulfillment of the Requirements for the Degree of Doktor der Naturwissenschaften

February, 2009

"Ich bin zwar nur ein Droschkengaul, -
doch philosophisch regsam;
der Fress-Sack hängt mir kaum ums Maul,
so werd ich überlegsam.
Ich schwenk ihn her, ich schwenk ihn hin,
und bei dem trauten Schwenken
geht mir so manches durch den Sinn,
woran nur Weise denken.

Ich bin zwar nur ein Droschkengaul, -
doch sann ich oft voll Sorgen,
wie ich den Hafer brächt' ins Maul,
der tief im Grund verborgen.
Ich schwenkte hoch, ich schwenkte tief,
bis mir die Ohren klangen.
Was dort in Nacht verschleiert schlief,
ich konnt' es nicht erlangen.

Ich bin zwar nur ein Droschkengaul, -
doch mag ich Trost nicht missen
und sage mir: So steht es faul
mit allem Erdenwissen;
es frisst im Weisheitsfuttersack
wohl jeglich Maul ein Weilchen,
doch nie erreicht's - oh Schabernack -
die letzten Bodenteilchen."

– Christian Morgenstern

– Dedicated to my nuclear family

Contents

1. Introduction 5
2. Principles of Genetic Association Studies 9
2.1. Genotypes 9
2.2. Shaping of genotype distributions 9
2.3. Phenotypes 15
2.4. Further aspects 16
3. Methods in statistical genetics 18
3.1. Linkage analysis 18
3.2. Association studies 19
3.3. The family based association test 23
3.4. Segregation analysis 24
3.5. Notes on predictive sampling strategies 28
3.6. Genotype-based methods 29
3.7. Practical aspects in association mapping 31
3.8. Wrapup 31
4. Notation and assumptions 32
4.1. Data 32
4.2. Assumptions used in the likelihood framework in this thesis 35
5. Likelihood framework 38
5.1. The Penetrance Model 38
5.2. Likelihood for a candidate region 39
5.3. Extensions 44
5.4. Parameter estimation 47
5.5. Statistical testing 48
6. Properties of the likelihood 49
6.1. Identifiability of conditionally decomposable likelihoods 49
6.2. Identifiability of likelihood L 50
6.3. Checking identifiability conditions for a concrete example 58
6.4. Consistency of MLEs 61
6.5. Identifiability of the random effects models 63
7. Bayesian approach 65
7.1. Data augmentation 65
7.2. Collapsing 66
7.3. Sampling strategy 67
7.4. Prior distributions and deterministic relationships 67
7.5. Densities for sampling distributions 68
7.6. Updating $H^\star$ 69
7.7. Updating of β 70
8. Simulation study 71
8.1. Parameter estimation and comparison of family structures 71
8.2. Parameter estimation under the null hypothesis 72
8.3. Ascertainment 72
8.4. Family structure 72
8.5. Haplotype analysis 75
8.6. Asymptotic normality of parameter estimates 75

8.7. Power comparison with the FBAT statistic 80
8.8. Random effects model 82
8.9. MCMC simulations 83
8.10. Computational issues 84
9. Alzheimer's disease 87
9.1. The ApoE locus 87
9.2. The ApoE/Alzheimer's data set 88
10. Discussion 96
10.1. Robustness 96
10.2. Guidance of experimental design 96
10.3. Ascertainment 97
10.4. Haplotype effects 97
10.5. Limitations 97
10.6. Biological relevance 98
10.7. Genome wide association scans 99
10.8. Future work 101
11. Acknowledgments 102
Appendix A. Abbreviations and Glossary 103
Appendix B. Haplotype reconstruction 105
B.1. C code for construction of $\mathcal{H}(G)$ in nuclear families 105
Appendix C. Appendix simulations 111
C.1. Implementation of the likelihood 111
C.2. Estimation of Fisher-Information 117
C.3. Data simulations for a single observed locus 122
C.4. Simulations for the two locus case 168
C.5. MCMC convergence plots 181
Appendix D. Alzheimer's data set 185
References 189

1. Introduction

Scope of this Thesis

Human genetics tries to elucidate how genetic variation explains variation in observable human traits. Recent developments have led to a change of paradigm in the analysis of complex traits, *i.e.* traits that do not follow Mendelian inheritance and occur commonly in the population (*e.g.* hypertension, diabetes, obesity, dementia). First, the whole genome is investigated to identify genetic regions that may be associated with disease outcome (genome wide step). In followup studies these regions are characterized more finely and often genetic models for a given region based on the given disease are built (fine mapping step) [41, 14, 65]. This thesis aims to improve the fine mapping of a genetic region based on family data.

Background

The sequencing of the human genome [53, 106] paved the way for genome wide analyses in complex disorders by allowing to characterize common genetic variation. Further developments were driven by the observation that many diseases are defined as the tail of normal phenotype distributions (*e.g.* blood pressure - hypertension, body weight - obesity, IQ - dementia, etc.). This supported the idea that the same common genetic variation influencing normal traits should also be causal in common disease [65] (common gene - common disease hypothesis).

Recent technological and conceptual developments make it now possible to analyze the whole genome in individuals with respect to common genetic variation [60, 75, 41, 14] (genotyping). In order to optimize the amount of genotyping, representative, so called tagging markers (or tagging SNPs, see below; see *e.g.* [9, 14]; tagging stage) are chosen to each represent a small genetic region. The first question is: Do the data support a genetic contribution to the disease? In typical studies, ca. 500.000 tagging SNPs are investigated and analyzed one by one. This imposes a multiple testing problem, which requires low p-values (p $\approx 10^{-7}$; genome wide significance) for individual tests to be deemed significant. General methods like Bonferroni-Holm correction or the false discovery rate are employed as well as methods exploiting information about the genetic setting such as correlations between tagging SNPs (see *e.g.* [89, 44]).

Another goal is to understand the underlying biology that causes disease. Therefore, the genome wide analysis is usually followed by a fine mapping step that investigates regions identified in the first step more closely. Fine mapping might include adding markers and replication of findings in an independent sample, sometimes based on a different study design (family based vs. case control). Often, the statistical focus is still on testing rather than on estimation or prediction. For example, the transmission disequilibrium test (TDT) and its extensions test for genetic association [91] (a recent review [104] lists close to 200 extensions). As these tests are robust against population stratification (see below), they sacrifice information that may otherwise improve genetic inference. This thesis introduces a genetic model that allows direct biologic interpretation. The frequency of a causal genetic variant at a marker (disease allele) that might be unobserved is estimated as well as its penetrance on disease, modeled by logistic regression. These parameter estimates can guide follow-up experiments in a direct manner. A full likelihood framework is presented, which separates this work from earlier related methods that also use latent genotypes (see [108, 107, 2]). The model can be applied to random samples of families as well as to families sampled on the basis of multiple affected members.

Approach

In this thesis, a latent class model for family based association studies is proposed. The model parameters are the joint distribution of observed markers and an unobserved true disease locus in a genomic region and a penetrance parameter measuring the impact of the putative disease allele on disease risk. An extension accounts for markers that are not linked (*i.e.* marker observations are independent) to the current region by modeling them via a random effect. First, a full likelihood setting of the model is presented and asymptotic properties are studied. Additionally, a Bayesian framework for the model is presented. In the Bayesian setting, *a-priori* knowledge that a given region is associated with a disease outcome can be incorporated. Model properties are assessed in simulations and the model is applied to an Alzheimer's data set.

Results

For the likelihood framework, identifiability is shown as well as consistency of parameter estimates. Results of the simulations show that parameters of interest can be precisely estimated with practically relevant sample sizes. Comparisons of different family structures show that the model is robust to variations in family structure, a result that is relevant for study design. To investigate how much power is sacrificed in a robust testing framework, a comparison study was conducted, showing big power advantages for the latent genotype model. This suggests, that the model considered in this thesis can also contribute towards gene identification in terms of power after the validity of assumptions has been checked. An application to an Alzheimer's dementia data set applies the methods to a real world problem. Results from this data set agree with prior findings. This is reassuring as the Alzheimer's genetics exceeds the complexity assumed in this model, thereby underlining a certain robustness to misspecification. Potentially, parameter estimates could have optimized the gene discovery phase in the ApoE region, had they been available.

Structure of the Thesis

In chapter 2 genetic principles as well as basic concepts in statistical genetics are explained, including a literature review of relevant previous work. Chapter 4 introduces the notation used in the statistical models and lists needed assumptions. The following chapter establishes a likelihood framework that is used throughout the thesis and gives several parametrizations thereof (chapter 5). In chapter 6 identifiability of the model parameters is addressed. Identifiability is verified for most sub-models and plausible arguments are given to encompass the entire model space. As the parameter space grows exponential with genetic complexity, a Bayesian approach is attractive and is explored in chapter 7. A simulation study (chapter 8) explores the properties of the frequentist likelihood as well as the Bayesian Markov Chain Monte Carlo (MCMC) for finite samples, including a power comparison with a well established method, the family based association test

(FBAT). The frequentist approach is applied to an Alzheimer's data set in chapter 9 and the thesis concludes with a discussion (chapter 10) highlighting weaknesses and strengths of the proposed methods and future extensions.

2. Principles of Genetic Association Studies

The main goal of genetic association studies is to predict phenotypes - observable traits of individuals - from genetic information. The aim of this chapter is to motivate the probability distributions involved, which in turn motivate the statistics being constructed later.

2.1. Genotypes.

On the molecular level, genetic information is encoded in DNA molecules - the chromosomes - which are composed of a sequence of four different chemical components called nucleotides. The nucleotides are labeled $\{A, C, T, G\}$ and stand for Adenin, Cytosin, Thymin and Guanin, respectively. The complete human genome (all chromosomes) is composed of 46 chromosomes. 44 of these chromosomes - the so called autosomes - occur in pairs of almost identical (homologous) chromosomes. One of these homologous chromosomes has been transmitted by the mother and the other one by the father. The remaining two chromosomes are called sex chromosomes and follow a homologous pattern in females (two X chromosomes), whereas in males a pair of chromosomes largely differing in size (X and Y) is formed (Figure 2.1).

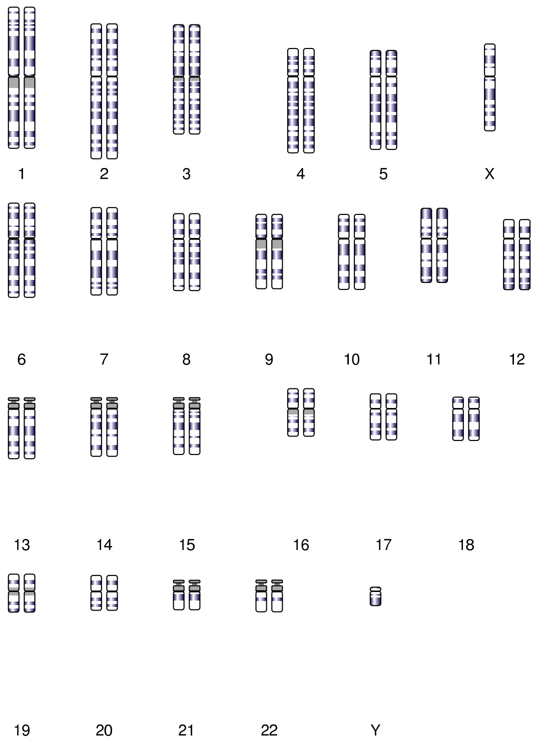
2.2. Shaping of genotype distributions.

2.2.1. *Genetic variation.*

Individuals are genetically unique because homologous chromosomes are not identical. On the population level, sequence differences occur at an average of 300-500 base pairs[1] (bp) [43]. The nature of these variations is categorized by the preassumed mechanism that produced the variation. The most common variations include single nucleotide exchanges (single nucleotide polymorphisms; SNPs), variation in copy numbers of short, repeated sequences (microsatellites) and insertions/deletions of short ($<$ 100 bp) sequences (table 2.1). The distinct variant sequences are called alleles. The

[1] The replication of genetic material is organized in such a way, that two DNA strands are tightly linked by chemical bonds. Both strands can be derived from each other by a simple bijective mapping ($f : \{A, C, T, G\} \rightarrow \{A, C, T, G\}, f(A) = T, f(G) = C$). The DNA molecule is therefore composed of two strands in which the base pairs A/T and G/C always face each other. In replication both strands serve as a template for a copy of the DNA molecule.

Figure 2.1: Diploid male caryogram



site is called locus and, conventionally, the variation is called a polymorphism if the smallest allele frequency is $\geq 1\%$. A generalization of alleles are haplotypes that denote the combination of adjacent alleles on a single chromosome (figure 2.2). Pairs of haplotypes are called diplotypes. Humans are thus genetically fully characterized by their diplotype distribution.

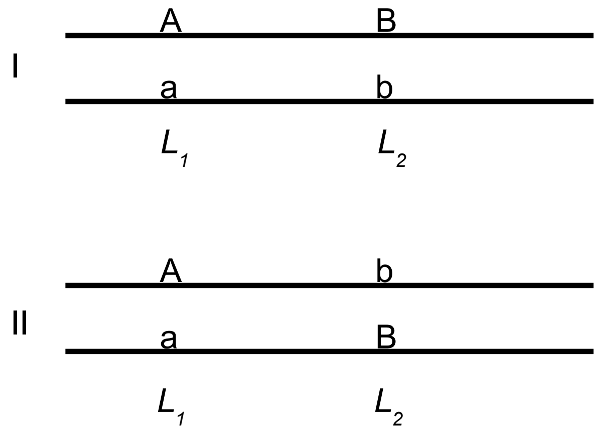
The models in this thesis are presented for SNPs as they are the most commonly used markers in large scale genetic studies. Affordable technical solutions for genotyping (the experimental process to observe genetic information) have been developed.

2.2.2. *Measurement of genetic variation - Haplotype ambiguity.* Most laboratory experiments to determine genetic information only consider one locus at a time and do not allow to determine which alleles lie on a single chromosome, known as phase information. This

Table 2.1: Different types of polymorphisms in the human genome. The column frequency gives this estimated absolute number of occurrence in the human genome [43]. These numbers are for polymorphisms for which the second most frequent allele has a frequency of at least 5%

Name (Synonyms)	**Description**	**Frequency**
Single nucleotide polymorphism (SNP)	Difference in a single base pair at a locus (*e.g.* bases 'A' and 'C')	$\sim 3 \times 10^6$
Microsatellites; variable number of tandem repeats (VNTR)	Variations in the repetition number of a small sequence motif (2 to 6 base pairs; *e.g.* 'CAG')	$\sim 10^5$
Insertions/Deletions	sequences either present or absent in an individual (*e.g.* retroposon derived)	$\sim 10^5$

Figure 2.2: Haplotype ambiguity. Cases I and II represent individuals for which alleles A/a at locus L_1 and B/b at locus L_2 are arranged in different combinations on chromosomes (represented by lines) such that separate and independent observation of L_1 and L_2 (genotypes) yields the same information in both cases.



	Haplotypes	Genotypes
I	AB — ab	Aa, Bb
II	Ab — aB	Aa, Bb

concept is illustrated in figure 2.2. Both diplotypes carry the same alleles at each locus, yet the configuration on the chromosomes cannot be determined through the genotypes. Observations of alleles without phase are called genotypic information. The process of obtaining genotypic information is called genotyping. In this thesis all observed genetic information is assumed to be unphased.

If the same allele is observed twice, the individual is said to be homozygous for that allele, otherwise the individual is said to be heterozygous[2].

2.2.3. *Likelihood for haplotypes.* In this thesis the two haplotypes forming the diplotype of an individual are assumed to be sampled independently and identically from the previous generation following a multinomial distribution. This imposes a parental origin on the diplotype: $H_o = (H_M, H_P)$ is composed of maternal haplotype H_M and paternal haplotype H_P. Let us further assume that we observe K bi-allelic loci for which 2^K possible haplotypes are observable (compare figure 2.2) denoted by numbers $1, ..., 2^K$. We can then express the likelihood for diplotype H_o as follows:

$$P(H_o = h) = P(H_M = h_M)P(H_P = h_P) = \eta_{h_M}\eta_{h_P}.$$

Here, $\eta_{h_M}\eta_{h_P}$ are the parameters of the multinomial distribution corresponding to haplotype frequencies. Firstly, we assume that parental origin cannot be determined. Formally we can express this fact by imposing a lexicographic ordering on haplotypes in diplotypes. If the set of diplotypes with parental origin is denoted by $\mathcal{H}_o = \{1, ..., 2^K\}^2$ and the lexicographically ordered set by $\mathcal{H}_u = \{(h_1, h_2) \in \mathcal{H}_o | h_1 < h_2\}$, the mapping $\pi_o : \mathcal{H}_o \rightarrow \mathcal{H}_u, (h_1, h_2) \rightarrow (h_{(1)}, h_{(2)})$ destroys information about parental origin (where $h_{(i)}$ is the order statistic). Then, the likelihood of diplotype $H_u = h_u = (h_1, h_2)$ without parental origin is therefore given by:

$$\begin{aligned} P(H_u = h_u) &= \begin{cases} \eta_{h_1}^2 & \text{for} \quad h_1 = h_2 \\ 2\eta_{h_1}\eta_{h_2} & \textit{else} \end{cases} \\ &= \sum_{(h_M, h_P) \in \pi_o^{-1}(h_u)} P(H_M = h_M)P(H_P = h_P), \end{aligned} \tag{2.1}$$

The right equation illustrates the fact, that loosing ordering with respect to parental origin is a projection which induces the probability on H_u. We will call these diplotypes unordered in short. Secondly, we assume that phase information is missing for genotype

[2] In practice, most experimental methods distinguish the situations where either a single allele or two alleles are observed. In the former case it is assumed that two identical alleles exist at the locus under investigation and the individual is homozygous and heterozygous in the latter case.

observations. If we represent a haplotype h_i by the individual alleles composing the haplotype $h_i = (a_{1i}, ..., a_{Ki})$, genotypes are given by the lexicographically ordered pairs of alleles at each locus[3]. For example, the genotypes corresponding to diplotype (h_1, h_2) is given by $g = ((a_{1(1)}, a_{1(2)}), ..., (a_{K(1)}, a_{K(2)})$ with $a_{i(1)} < a_{i(2)}$ for $i = 1, ..., K$. Therefore, $g \in \mathcal{G} = \{(1,1),(1,2),(2,2)\}^K$. We denote this mapping $\pi_g : \mathcal{H}_u \to \mathcal{G}$ and arrive at the following likelihood for genotypes:

$$P(G = g) = \sum_{(h_M, h_P) \in \pi_o^{-1} \circ \pi_g^{-1}(g)} P(H_M = h_M) P(H_P = h_P),$$

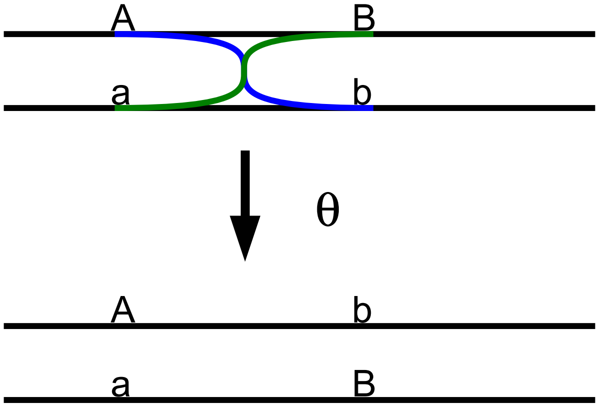
This recipe allows to state likelihood functions depending on observed genotypes in terms of diplotypes with known parental origin. Identifiability issues will be discussed later. As the likelihood for unordered diplotypes is only slightly more complex than the likelihood on $\mathcal{H}_o$ but is computationally more efficient, we consider unordered diplotypes in the following and use the notational convention $h_1 | h_2 = g$ to express $(h_1, h_2) \in \pi_g^{-1}(g)$.

2.2.4. *Mutations.* Each sequence variation is introduced into a population by a mutation event in a single individual. In a population of finite, constant size the allele frequency follows a random process, when alleles of the next generation are sampled randomly and independently from the previous generation (genetic drift). Therefore, the new allele eventually replaces the original one (fixation) or is lost again. As a polymorphism, by definition, has allele frequencies $\geq 1\%$ for all alleles, these have to have been present in the population for a certain number of generations since their appearance, which started with a frequency of $1/(2N)$, when N denotes the population size [40, 45].

Another mechanism that changes allele frequencies is selection, which implies non-random sampling of alleles from one generation to the next, which can more rapidly lead to the fixation of an allele.

[3] Assuming some numbering of alleles, say $a_{ij} \in \{1, 2\}$

Figure 2.3: Recombination in gametes



2.2.5. *Recombination.* The joint distribution of alleles is also influenced by a process called recombination. In progenitor cells of gametes (eggs in females, sperms in males) homologous chromosomes pair with each other, and for two given loci a break point is formed between these loci with probability θ, the recombination fraction, which depends on the specific pair of loci. Figure 2.3 illustrates that haplotypes $(A, B), (a, b)$ occur in the parent cell, haplotypes $(A, b), (a, B)$ emerge in daughter cells. Through recombination haplotype distributions converge to the product distribution of genotypes. However, mutations usually introduce new dependencies between alleles. Although recombination events do not occur uniformly across the genome, a recombination fraction of 1% (1 centimorgan) is roughly equated to 10^6 base pairs [53].

2.2.6. *Linkage disequilibrium.* The haplotype distribution at any set of loci can be represented by a multinomial distribution. At two loci this distribution can be re-parametrized by specifying allele frequencies and pairwise covariances between alleles at the two markers. This pairwise covariance is also referred to as linkage disequilibrium (LD). If both loci are bi-allelic, for example locus one has alleles A, a and locus two has alleles B, b, LD is the covariance $\delta = \mathrm{Cov}(X, Y)$. Here, X and Y are indicator variables $X = I\{\text{A observed at locus one}\}$ and $Y = I\{\text{B observed at locus two}\}$. The triple $(p_1, p_2, \delta) = (P(X = 1), P(Y = 1), \mathrm{Cov}(X, Y))$ characterizes the joint distribution at both loci. If recombination events can be assumed to be independent from haplotypes, the haplotype distribution tends to the product distribution of alleles at single loci (and therefore $\delta \to 0$). Therefore, for larger recombination fractions θ a smaller δ is expected [99]. However, the actual LD pattern is influenced by other mechanisms, like mutations.

2.2.7. *Identical by State (IBS) vs Identical by Descent (IBD).* Related individuals are dependent, as stretches of DNA are passed along from generation to generation. Consider the observation of two alleles X, Y, each from a different individual. If $X = Y$, the alleles are said to be identical by state (IBS). If the two individuals are related, and X and Y are known to derive from a common ancestor, X and Y are said to be identical by descent (IBD). Consider, for example, a parent with genotype AA and an offspring with genotype Aa. One parental allele A is then shared IBD between the individuals, since one A allele has to be transmitted from parent to offspring. Alleles shared IBD also share neighboring alleles, *i.e.* haplotypes, since chromosomes are passed along as a whole and only recombinations[4] limit the extent of haplotypes shared by IBD alleles. In conclusion, alleles shared IBD share larger surrounding haplotypes than IBS alleles.

2.3. **Phenotypes.** Phenotypes Y are observable traits in individuals, including binary traits, like disease status, continuous traits, like body height or blood pressure, and multivariate traits, like facial characteristics. This thesis focuses on binary traits.

2.3.1. *Penetrance function.* The penetrance function formalizes the relationship between genotypes G and phenotypes Y by specifying the conditional distribution $Y|G$. For a binary phenotype $Y \in \{0, 1\}$, a logistic distribution can be assumed:

$$P(Y = 1|G) = \frac{\exp(\mu + X(G)\beta)}{1 + \exp(\mu + X(G)\beta)}.$$

Here μ denotes the baseline penetrance, and $X(G)$ denotes a score for genotype G, which is determined by the so-called mode of inheritance. For example, let us assume that

[4] Recombinations are the most frequent event to alter chromosomes, and are therefore the most relevant factor to modify haplotypes

alleles a and A can be observed, thus genotypes $\{aa, aA, AA\}$ can occur, which might be enumerated with $0, 1, 2$, counting allele A.

Table 2.2: Score functions for modes of inheritance

Mode of inheritance	X(0)	X(1)	X(2)
Dominant	0	1	1
Recessive	0	0	1
Additive/dose model	0	1	2

In the dominant disease model, carrying at least one allele A changes disease probability from the baseline; the recessive model requires two alleles A to be present in a genotype to change disease probability from baseline.

2.4. **Further aspects.**

2.4.1. *Hardy-Weinberg equilibrium and population structure.* A common assumption, also used in this thesis, is that the genotype distribution at a given bi-allelic locus can be parametrized by a single parameter, namely the allele frequency of one allele. If both alleles are picked independently from the parental generation and allele A has frequency p, genotypes AA, Aa, aa have frequencies $p^2, 2p(1-p), (1-p)^2$, respectively. The assumption of independent random transmission of alleles in a population is called Hardy-Weinberg equilibrium (HWE). For large populations with random mating, *i.e.* a cygote is formed from a random pick of a sperm and an egg, HWE has been shown to hold [92].

However, in certain scenarios HWE might be violated, for example when a population is a mixture of several sub-populations with differing allele/haplotype distributions (population stratification). Several proposals have been made to deal with population structure [16, 3, 79, 80, 83]. One framework [79, 80] allows to estimate membership probabilities of sub-populations for an individual. These estimates can be used to stratify likelihoods for sub-populations and weigh the observations according to their membership probabilities [79]. In this thesis, however, we assume an underlying homogeneous population.

2.4.2. *Complex diseases.* Monogenic diseases are characterized by strong effects of a single gene, typically resulting in many affected individuals per family. Such situations can be explored by linkage analysis which is reviewed in section 3.1. By contrast, many diseases do not exhibit strong familial aggregation patterns (sporadic occurrence). Predisposition to these diseases, which include cardiovascular diseases (heart attack, stroke, atherosclerosis etc.), autoimmune diseases (multiple sclerosis, rheumatoid arthritis, inflammatory bowl disease etc.), metabolic disorders (obesity, diabetes) and psychiatric disorders (schizophrenia) among others, is thought to be caused by many genes with potentially small effects (polygenic) and additional environmental factors. Complex diseases are likely to be influenced by genetic factors in the same way as are normal traits (*e.g.* [74]). Both are considered to be influenced by common alleles [65]. These alleles can not have strong effects, since otherwise fixation should have occurred (compare section 2.2.1). The similarity of complex diseases and normal traits is also reflected by some definitions of phenotypes. For example, blood pressure distribution is used to define "hypertension", with somewhat arbitrary cut-off values.

It follows that complex diseases should be studied by analyzing polymorphisms that have been characterized in random samples as has been done in the International Hapmap Project (section 2.2.1) [13, 14]. The following paragraphs summarize methods used to characterize genetic distributions (haplotype reconstruction), phenotype distributions (segregation analysis) and association of polymorphisms with phenotypes. Also some essential differences between family based studies and case control studies are highlighted: families harbor more information than case control studies that can either be used to make more robust inference or to glean genuinely new information such as information about unobserved loci as in this thesis.

3. Methods in statistical genetics

Experimental as well as statistical methods have undergone rapid development in recent years. Initially, few markers in the genome were available for genetic analysis. This dictated that the initial focus was on monogenic diseases, diseases that strongly aggregate in families and show transmission from generation to generation. The recombination rate θ between an observed marker and the unknown disease locus can be estimated by assuming a penetrance function for the disease locus, using likelihood based methods. This type of analysis, known as linkage analysis, evaluates co-segregation of marker alleles and disease status, and can be conducted with relatively few markers. It is described in more detail below (section 3.1). As genes for monogenic diseases were successfully mapped (located) and more markers became available, case-control and family based association designs have become increasingly popular. Association designs rely on correlations between disease variants and observed markers. Recent efforts have characterized the LD structure of the human genome (physical [22, 1, 13]; by simulations [48, 97]). In the following, two major approaches, linkage and association studies are discussed. Other strategies, like admixture mapping, are reviewed elsewhere (e.g. [31]).

3.1. **Linkage analysis.** Linkage analysis has been successfully employed in finding genes in monogenic disorders, *i.e.* conditions for which variation in a single locus causes disease [68]. The locus is often called disease locus and the allele present in affected individuals is called disease mutation. These conditions follow a well-defined mode of inheritance, *i.e.* dominant, recessive or additive. As detailed in section 2.2.5, any allele has a 50% chance of being passed on into a zygote. This implies that, for example, in dominantly inherited traits many family members are typically affected if such an allele segregates in a family. The recombination rate θ between an unobserved and the marker locus is estimated by using either a likelihood based approach or by comparing expected and observed numbers of IBD sharing, based on the idea that if a locus is closely linked to a disease locus (*i.e.* the recombination rate is low) the same allele should segregate over generations, together with the phenotype [52, 49, 50, 90, 11]. However, if the penetrance is low, or many genes are involved in a disorder, families with several affected members are hard to find. The minimum requirement on a family for linkage analysis are two affected individuals, and the affected sib pair (ASP) analysis is one of the standard designs (e.g. [6, 20]). However, it has been noted that linkage designs are not powerful if several loci contribute to the disease [84]. Also, linkage analysis has limitations to resolve between closely linked markers, since recombinations between them become rare (for estimates see e.g. [53, 66]). The typical resolution of linkage analysis is in the order of megabases (Mbs), as compared to kilobases in association mapping.

3.1.1. *Positional cloning.* If linkage analysis demonstrates that a marker is linked to a disease gene, more analyses are typically required for fine-mapping, as large stretches of DNA might be linked with the disease gene. Such regions can span up to several Mbs. To characterize the DNA sequence in the region and to search for changes that can be demonstrated to co-segregate with disease and suggest a plausible disease mechanism, a laboratory-based step called positional cloning is applied[5]. The human genome project [54, 106] has eased this process substantially, since it provides full sequence information for most regions in the genome, and sequences can be selected based on database searches for further investigation.

3.2. **Association studies.** Association studies consider alleles as exposures, and a disease status as outcome. They can be divided into case control studies using independent cases and controls and family based studies. Association studies in human genetics are conducted using a wide variety of both principle study designs and statistical models. In this chapter we are going to review both aspects of association studies:

- Case control studies with *i.i.d.* samples
- Family based studies with independent families.

Based on the phase uncertainty in genotyping (*c.f.* section 2.2.3), methods differ in how haplotypes are modeled. A possible categorization is:

[5] Cloning, in this context, means unspecific amplification of DNA sequences in bacteria, which can be followed up by sequence analysis to identify disease mutations.

- Two stage models estimating haplotype frequencies in a first step and using these in further analysis
- Likelihood-based approaches with simultaneous haplotype frequency estimation
- Genotype-based methods that only implicitly heed haplotype structure

3.2.1. *Case control studies.* Case control studies exploit the fact that recruiting affecteds for some binary phenotype should enrich causative alleles in this group as compared to a group sampled without ascertainment. Odds ratios of allele effects are usually much smaller in these studies than in linkage studies (OR $\approx 1.3 - 3$). Case control studies with independent cases and controls have been shown to have better power than either family based association or linkage studies [84, 100, 85], if enough markers are considered [105]. However, they are more sensitive to confounding (population stratification) [91] and do not allow for the identification of several interesting parameters (see below). Case-control studies have been used to successfully find disease genes, e.g. macular degeneration [46, 29, 17] and obesity [30].

3.2.2. *Family based association studies.* Family based association studies usually sample nuclear families, comprised of parents and offspring. Often these families are not random samples, but are ascertained on the basis of having at least one affected offspring. Sampling nuclear families allows one to condition on parental genotypes and therefore does not require assumptions on the parental genotype distribution (like HWE). This idea has been exploited in the Haplotype Relative Risk (HRR) [87] statistic and the Transmission Disequilibrium Test (TDT) [91]. The concept of making robust inference, without assumptions about the haplotype distribution has been formalized in the FBAT (family based association tests) tests [82, 51]. The idea is to condition on the minimal sufficient statistic that characterizes the null hypothesis. In the case of nuclear families with known parental genotypes the sufficient statistic consists of all parental genotypes and all phenotypes in the pedigree. Whittemore proposed an approach to estimate an association parameter with minimal variance that quantifies the strength of association between marker and phenotype based on family data [107]. Consistency of the parameter estimate is achieved by solving an estimation equation that is constrained in such a way that it is independent of the parental genotype distribution.

If distributional assumptions about the parental genotypes are made, a full likelihood approach allows for the estimation of association parameters that characterize the disease locus and the penetrance function. We explore this idea in this thesis.

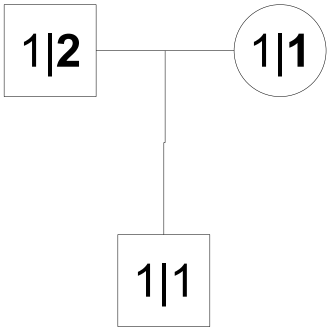
3.2.3. *Robust family based tests.* I want to mention the work by Göring and Terwilliger (2000) [26, 27, 25, 24] in some detail, as it closely relates to the model proposed in this thesis. They propose to decompose the likelihood $P(Y, G)$ of phenotypes and genotypes in a family as $P(Y, G) = \sum_{g^\star} P(Y|g^\star)P(g^\star|G)P(G)$. Here, g^* iterates all joint genotype combinations in the family at a putative disease locus. The likelihood contribution $P(g^\star|G)$ is modeled as a function of LD (δ), recombination fraction (θ), and allele frequencies. Profile likelihoods are used to eliminate either one of the parameters. One important difference to the model considered in this thesis is that pseudo-genotypes are assigned to the disease loci based on phenotypes and family structure. This assignment ensures the maintenance of alpha levels under the null hypothesis. Interpretation of parameter estimates, however, depend on this assignment. The main application of this model therefore is testing.

3.2.4. *The transmission disequilibrium test (TDT).* The TDT is a test for genetic association and followse the null distribution even when genotype or haplotype distributions do not follow HWE [91]. The main principle is to consider the distribution of offspring alleles, conditional on parental genotypes. Any parental allele has a chance of $\frac{1}{2}$ of being transmitted or non-transmitted such that the test statistic does not depend on marginal genotype/haplotype distributions. The concept is illustrated in Fig. 3.4. For each of the alleles of an offspring it is determined which parent transmitted that allele which implies that the other allele of this parent was not transmitted. This primary statistic results in a transmission table. Table 3.3 shows the simplest case for nuclear families with one offspring, and a single SNP marker, for which each parental transmission is counted in one of the cells.

Table 3.3: Transmission table in a nuclear family for alleles A, a. n_{xy} denotes the number of transmissions corresponding to parent-offspring pairs for which allele x was transmitted and allele y was not.

	transmitted	
	A	a
untransmitted A	n_{AA}	n_{aA}
a	n_{Aa}	n_{aa}

Figure 3.4: Transmission pattern in a nuclear family. Non-transmitted alleles are depicted in bold face.



Each parent transmits exactly one allele and does not transmit the remaining allele. Each nuclear family therefore contributes two entries in the table, and in case of no association, the table entries will be symmetric. When both parents and the offspring have identical heterozygous genotypes, both contributions lie symmetrically in off-diagonal elements, irrespective of the parental origin of alleles which cannot be determined in this case. Only if a sex specific direction of transmission is of interest, such observations are uninformative [111]. Some families are always uninformative, such as a nuclear family with only homozygous individuals.

The test statistic is given by

$$T = \frac{(n_{Aa} - n_{aA})^2}{n_{Aa} + n_{aA}},$$

the distribution of which converges weakly to a χ_1^2 distribution under the null hypothesis that there is no linkage or association of observed markers with disease.

3.2.5. *Extensions of the transmission disequilibrium test.* The TDT has been extended to accommodate, both, for missing data and for sibships, *i.e.* more than one offspring per nuclear family [55, 47, 38]. Also, the re-use of samples from linkage studies and X-chromosomal inheritance have been studied [63, 12].

3.2.6. *Biases in the transmission disequilibrium test.* It has been shown that bias can occur using the above extensions in data sets with both missing parental genotypes and sibships [63]. The same paper proposes a correction to obtain an unbiased test statistic. A second possible bias in the TDT results from segregation distortion, *i.e.* a deviation from Mendelian segregation at a locus. This has been shown to be present in human populations [112] and can only be avoided by testing for segregation distortion in an independent, population based sample.

3.3. **The family based association test.** The Family Based Association Test (FBAT) generalizes and formalizes ideas of the TDT. The null hypothesis in the FBAT is that there is no association or no linkage of observed markers with disease [82, 81]. In contrast to the TDT, FBAT can handle missing data without reintroducing HWE assumptions. While the FBAT does not require HWE and is robust to population stratification, parental phenotype information is not used - like with the TDT. For a binary trait the FBAT test statistic in a family with offspring genotypes $G_1, ..., G_n$ is

$$T = \sum_i \{X_i - E(X_i)\}$$

(see *e.g.* [37]), where X_i is a genotype score for genotype G_i of offspring i, similar to the score X^* defined later in this thesis. The expectation $E(X_i)$ is taken conditional on parental genotypes, and thus the marginal distribution of parental genotypes does not contribute to the statistic. The score X_i can be chosen such to reflect modes of inheritance. The test statistic cannot be expressed in closed form in general and is computed by an algorithm that isolates appropriate outcomes under missingness that contribute to the test statistic. For complete data, FBAT is identical to the TDT.

3.3.1. *Testing for association.* As opposed to the approach of this thesis the TDT and FBAT can only test for significant association of genetic markers with a trait. The aim of this thesis is to estimate parameters that characterize the genetic region and the genetic effects in addition to offering a testing framework. However, this comes at the cost of additional assumption (see next chapter), *e.g.* TDT/FBAT do not require Hardy-Weinberg equilibrium assumptions.

3.4. **Segregation analysis.** In contrast to linkage or association studies, segregation analysis is a statistical framework to exploit information from phenotypes only in nuclear or more complex family structures without genetic information [2]. The phenotypes give indirect information about underlying genotypes, and their relationship is modeled by a penetrance function. The likelihood of segregation analysis for a single family is given as follows:

$$L(Y) = \sum_{g^\star} L(Y, G^\star = g^\star) = \sum_{g^\star} L(Y|G^\star = g^\star) L(G^\star = g^\star)$$

Here Y is the phenotype vector and $g^\star$ iterates all genotype combination at the disease locus. Founding individuals, i.e. persons without parents in the pedigree, are assumed to be a random sample from the population. The similarity with the likelihood considered in this thesis occurs under the null hypothesis of no association between observed genotypes and phenotypes. In this instance, the likelihood in this thesis can be factorized into separate terms for phenotypes and genotypes (see section 5.2.4). The contribution of the phenotypes is the same as the likelihood in segregation analysis. Thereby the likelihood component is given as in sec. 2.4.1, i.e. a single parameter p_A parametrizes the allele frequency at a SNP locus. For non-founders $L(G^\star_i = g^\star_i|G^\star = g^\star) = L(G^\star_i = g^\star_i|G^\star_{iM} = g^\star_{iM}, G^\star_{iP} = g^\star_{iP})$, i.e. the likelihood only depends on parental genotypes $(g^\star_{iM}, g^\star_{iP})$.

Commonly, the expectation of the penetrance function in segregation analysis $L(Y|G^\star = g^\star)$ is assumed to be a linear function of several effects on the inverse scale of some link function h:

$$E(h^{-1}(y_i|g_i^\star = j)) = \mu_j + \nu_i + \epsilon_i.$$

μ_j is a mean associated with genotype $g_i^\star = j$, ν_i is a random polygenic effect, *i.e.* an effect modeling effects of multiple minor alleles, assumed to be normally distributed with mean zero and unknown variance σ_ν, and ϵ_i is assumed to be a random, normal environmental effect with mean zero and unknown variance σ_ϵ that is independent for different family members. The link function h is the identity for normally distributed traits and the logistic function for binary traits. The random effects are considered to be independent of one another. The phenotypes in the family therefore have a multivariate normal distribution and the polygenic effect ν_i induces dependency by means of relatedness of individuals (analogous to sec. 5.3.2).

The parameters $\theta = (\mu, \sigma_\nu, \sigma_\epsilon)$ can be estimated via maximum likelihood. The model can be extended to account for covariates by including an additive term βx_i into the penetrance function (x_i being the covariates and β is a parameter modeling the effect). One major goal of segregation analysis is to determine the mode of inheritance which is reflected in the parameter vector μ. To allow for a model comparison, nested models have to be established. One way is to compare genetic models to a model that includes only environmental effects, thereby making phenotypes independent. Extensions that parametrize the transmission probabilities have been proposed [2].

3.4.1. *Two stage haplotype based analysis.* In a first step phase uncertainty is resolved and in a second step standard analysis is conducted, assuming known haplotypes. The first step involves calculating haplotype probabilities for individuals. Haplotype reconstruction has been demonstrated to be highly accurate [62].

3.4.2. *EM algorithm for haplotype frequency estimation.* The EM algorithm was among the first methods proposed to conduct analyses on a haplotype basis [21]. Using a standard genetic model assuming HWE (*c.f.* section 2.4.1), and given known haplotype frequencies, it is straightforward to establish a likelihood for genotype data, as was shown in section 2.2.3:

$$P(G = g) = \sum_{(h_M, h_P) \in \pi_o^{-1} \circ \pi_g^{-1}(g)} P(H_M = h_M) P(H_P = h_P).$$

A likelihood for genotypes given parameters for the haplotype distribution is then given by:

$$\begin{aligned} & L(\theta = (\eta_1, ..., \eta_m); G = (G_1, ..., G_n)) \\ = & \sum_{h_1 | h_2 = g} P_\theta(h_1, h_2). \end{aligned}$$

Given the likelihood (3.2), the EM algorithm is established as follows: (1) The E step involves estimating genotype frequencies, given known (initial) haplotype frequencies, (2) the maximization step maximizes the likelihood for haplotype frequencies, given genotype frequencies:

$$\hat{\eta}_h = \sum_{i=1}^{n} \left(\sum_{h_1 | h_2 = g} \frac{1}{2} P_\theta(h_1, h_2)(I\{h = h_1\} + I\{h = h_2\}) \right), \tag{3.2}$$

with $P_\theta(h_1, h_2)$ from section (2.2.3). $I\{h = h_1\} + I\{h = h_2\}$ counts the number of h haplotypes in diplotype (h_1, h_2). $\hat{\eta}_h$ is therefore the weighted count of possible occurrences of haplotype h in the observed genotypes with the weights being genotype frequencies from the expectation step.

3.4.3. *A Bayesian approach to haplotype reconstruction: PHASE.* In more recent work, haplotype frequency estimation has been conducted using Bayesian models [95]. The basic idea is, that for a given genotype haplotypes are predicted using haplotype predictions for all other genotypes. The ensuing Markov chain is of the form $P(H^{(k+1)} | H^{(k)})$. More specifically the reconstruction $H^{(k+1)}$ differs from $H^{(k)}$ for only a randomly chosen individual. The haplotypes for this individual, say $H_i^{(k+1)}$ are inferred from the remaining haplotypes, say $H_{-i}^{(k+1)}$. Therefore the probability $P(H_i^{(k+1)}|H_{-i}^{(k+1)})$ specifies the Markov chain, which can be restated as $P(h_1|H_{-i}^{(k+1)})P(h_2|H_{-i}^{(k+1)}, h_1)$ if $H_i^{(k+1)} = \{h_1, h_2\}$. It remains to specify the haplotype distribution of single haplotypes; given another set of haplotypes. The authors propose to use a predictive distribution of the following form:

$$P(h|H,G) = \sum_{a\in A}\sum_{s=0}^{\infty} \frac{r_a}{r}\left(\frac{\theta}{r+\theta}\right)^s \frac{r}{r+\theta}(P^s)_{ah}, \tag{3.3}$$

with A being the set of haplotypes in some finite population, $r = |H|$ the cardinality of H, $r_a = |\{h \in H|h = a\}|$ the number of a haplotypes in H, θ is a mutation parameter and P is a matrix of transition probabilities between members of A. s is drawn from a geometric distribution. The motivation for this distribution is given by reference to an evolutionary model. This distribution approximates the probability that h is sampled from the (large) population H given that a neutral coalescent process holds [93]. The neutral coalescent process assumes that a population has discrete transitions between generations of constant size for which each haplotype for a new generation is sampled uniformly and independently from the previous generation. Additionally, random mutation events are superimposed on the new haplotypes as specified by a transition matrix (the P from above) and a mutation rate (θ). A neutral coalescent is a common way to simulate genotype data and has been successfully employed to predict certain properties of the genome [1]. It can therefore be expected that data simulated from a neutral coalescent process should be well analyzed by the method. Indeed the authors show that their Bayesian method outperforms the EM algorithm and variants thereof as measured by the MSE and other criteria [94]. Apart from being more accurate in certain situations, the PHASE method has the advantage that it is possible to include more loci into the analysis than for the EM algorithm, as the set of reconstructed haplotypes H remains more manageable.

3.4.4. *A divide-and-conquer strategy for predictive sampling.* Haplotype reconstructions involve the computation of sets of compatible haplotypes for given genotypes. The cardinality of these sets grows exponentially with the number of analyzed loci (*c.f.* section 2.2.1). It has therefore been proposed to use divide-and-conquer strategies in haplotype reconstruction [72] (PL algorithm). These strategies involve two stages. The first step requires a partitioning of the set of observed loci. Then a haplotype reconstruction method is applied to the subsets individually. Any method to estimate haplotype frequencies can be used. The second step now involves combination of these individual results. Also there are several possibilities to conduct step two. The authors consider

- Linear ligation of adjacent sets
- Hierarchical combination of sets, doubling the size of sets in each step.

For this divide-and-conquer strategy, it is not possible to use the identical predictive distribution as is used in PHASE, since a similarity measure on complete haplotypes is used in the predictive distribution there. Instead, haplotypes are assumed to be multinomially distributed, with a Dirichlet prior. The resulting predictive distribution is:

$$P(h_i^\star | H_{(-i)}^\star, G) = P((h_{i1}^\star, h_{i2}^\star) | H_{(-i)}^\star, G) \propto (n_1 + \gamma_1)(n_2 + \gamma_2).$$

Here, $\gamma_1, ..., \gamma_M$ are parameters of the Dirichlet prior and are chosen to be uninformative.

3.5. **Notes on predictive sampling strategies.** By making more assumptions, the PHASE algorithm performs better than the divide-and-conquer strategy [72] at the cost of being computationally more involved. Both Bayesian approaches seem to perform better than the EM-algorithm [94, 62]. This can be explained by the additional assumptions made and partly by the fact, that both methods condition on observed genotypes. This fixes allele frequencies in each updating step and thereby does not account for variability in allele frequencies in contrast with the EM-algorithm. This is demonstrated later in this thesis (section 7.6). In practical applications the main goal is often to resolve phase (*i.e.* infer diplotypes from genotypes) for individuals and deriving individual posterior distributions, a task well fulfilled by the Bayesian methods. However, in their current implementations, they fail to account for the variability in marginal allele frequencies, as they condition on observed genotypes in contrast with the EM algorithm.

3.6. **Genotype-based methods.** Models can also be based on genotypes rather than haplotypes. One potential problem is that of parameter dimensionality. For n SNPs 2^n potential haplotypes exist compared with 3^n genotypes.

3.6.1. *Using global test hypotheses.* All methods reviewed here choose all subsets of cardinality L from n loci under scrutiny, giving rise to a number of $\binom{n}{L}$ analyses. Non-parametric tests can be used to evaluate differences between a case and a control group for each of the subsets [7], for example by counting genotype combinations at the different loci and conducting a χ^2 of independence for a case control data. Since the tests are not independent for subsets with overlapping loci, resampling techniques can be used to test a combined test statistic over all subsets, for example the sum of (normalized) χ^2-statistics. Genetic models can be incorporated by assigning scores to genotype combinations. Cell counts can then be computed by computing $|\{g_i|I_A(g_i) = 1\}|$ in each group. Here I_A is an indicator on scored genotypes. In general the table looks like:

	A	$\bar{A}$
Y = 0	n_{11}	n_{12}
Y = 1	n_{21}	n_{22}

with Y indicating the phenotype and A being the occurrence of some allele combination (AC) or genotype.

Global significance tests can be combined with graphical methods to highlight interactions between loci.

3.6.2. *Set association.* A similar approach tries to isolate subsets of alleles for which an association with a phenotype can be found [34]. The null hypothesis here comprises only a certain AC that is determined in a preceding step using a heuristic search through possible ACs. The search step involves the evaluation of the association of single alleles. Two χ^2 statistics are used: (1) a contingency table test as in sec. 3.6.1; (2) a χ^2 goodness-of-fit test for HWE is applied. The final statistic is the product of the two χ^2 statistics. Subsets of ranked statistics are considered, where the cardinality of subsets increases in a stepwise procedure. The significance level of each subset is evaluated using a permutation test randomly changing disease status in a case control sample, and the minimum p-value from this procedure is chosen to indicate the best associated AC. A global significance level is then computed using again a permutation test. One problem with this procedure is that the sequence of test statistics has to be stopped at a certain point, say N, which has to be chosen arbitrarily. On the other hand, N influences the final permutation step and therefore is subject to heuristic choices.

3.6.3. *Combinatorial choice of genotypes.* Another application of genotype based methods considers systematic combinations of alleles or genotypes similar to the approach in section 3.6.1. This approach was developed for a quantitative trait [71]. All genotype combinations are iterated for a fixed number of m loci. The procedure can be broken down into three steps:

- Group genotype combinations into disjoint sets (partitions) based on some similarity measure and evaluate partitions using an ANOVA model.
- Use cross validation methods to evaluate the performance of partitions.

Grouping of genotypes in the first step reduces the complexity of the genetic model, since only indicator variables on the partitions are used as factors in the ANOVA model. Sparse partitions can therefore reduce the number of degrees of freedom. Partitions are evaluated on the basis of how much of the variance of the phenotype is explained by the partition factor and a heuristic cutoff is used to include partitions into the model. The second step evaluates the predictive power of the models that were selected in the first step. The cross-validation procedure estimates mean parameters for a training set chosen from the original data and the sum of squares is computed for the observed and predicted values for the remaining test set. The last step reduces the set of partitions selected in the first two steps into a smaller set by taking into account sums of squares computed in step 2. This method does not yield rigorous p-values but relies on the cross validation procedure to ensure validity.

3.6.4. *Multifactor-dimensionality reduction.* A closely related work to the combinatorial choice of genotypes is multifactor-dimensionality reduction (MDR) [86], which uses machine learning techniques to find predictors of phenotypes based on genotype combinations. The search through model space is guided by a step-wise procedure that relies on cross-validation for verifictation.

3.7. **Practical aspects in association mapping.** The HapMap [13] is an ongoing effort to physically characterize the LD structure in human populations. Several million SNPs are expected to exist in most human populations. So called candidate gene studies select SNPs on the basis of functional knowledge about genes with a plausible involvement in the phenotypes. In contrast, all SNPs in the genome can be analyzed in a whole genome analysis (WGA). Because of redundancies based on LD patterns, it is possible to select a subset of SNPs (tagging SNPs) for genotyping (*e.g.* [4]) in both approaches. This tagging SNP selection can be based on the HapMap data. However, the small sample size ($N \sim 40$) estimates have a large variance [101, 102].

3.8. **Wrapup.** Previous approaches to identifying genes involved in complex disorders have mostly been concerned with hypothesis testing (see section 3). In this thesis, an approach to estimating the joint distribution of observed alleles and a possibly unobserved disease allele is estimated, together with parameters in the penetrance function, one potential benefit being, that the experimental design ensuing localization of significantly associated alleles can be guided by parameter estimates.

4. Notation and assumptions

4.1. **Data.** We assume that N nuclear families are sampled into our study, with n_i denoting the family size of family i. The phenotypes of the ith family are $\mathbf{Y_i} = (Y_{i1}, ..., Y_{in_i})$. Indices 1 and 2 designate parents and indices $3, ..., n_i$ denote offspring. The phenotype is a binary trait, that is $Y_{ij} = 1$ for affected and $Y_{ij} = 0$ for unaffected individuals.

We assume the true disease locus is unobserved, and K marker loci in the same region as the disease locus are measured for each individual. Such a region would typically span a range of 50-100kb, depending on the population and the specific location in the genome [41]. We denote the observed multi-locus genotype of individual j in the ith family by $G_{ij} = (G_{ij1}, ..., G_{ijK})$. The observed joint genotypes for the ith nuclear family are $\mathbf{G_i} = (G_{i1}, ..., G_{in_i})$.

The observed data for each family thus consist of phenotypes and genotypes at the observed loci, $(\mathbf{Y_i}, \mathbf{G_i}) = (Y_{i1}, ..., Y_{in_i}, G_{i1}, ..., G_{in_i})$.

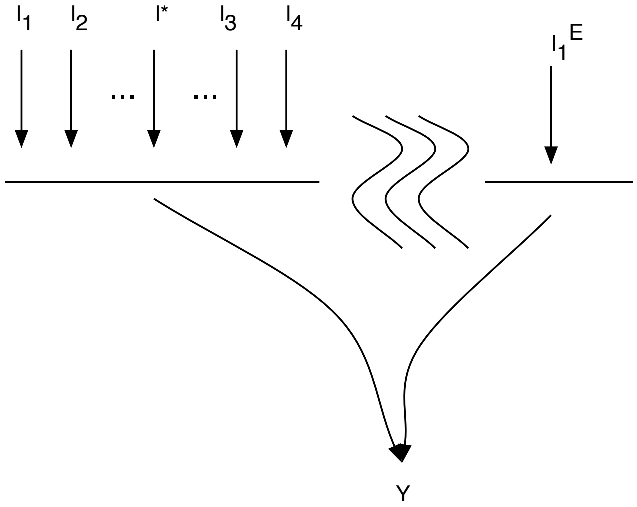
4.1.1. *Notation for haplotypes.* Figure 4.1 shows the assumed genetic model. In a candidate region K loci $L = \{l_1, ..., l_K\}$ are observed (in the example $K = 4$). A disease locus $l^\star$ is located in the region and further loci influencing the disease might be located in unlinked regions in linkage equilibrium, the set of which is denoted with $L_E = \{l_{E1}, ..., l_{EM}\}$. The set of all loci in the model is therefore given by:

$$\mathcal{L} = L \cup L^\star \cup L_E,$$

with $L^\star = \{l^\star\}$. L_E is modeled by a random effect, whereas $L \cup L^\star$ are directly included in the model. We denote with $\mathbf{G_i}^* = \mathbf{g_i}^*$ the joint observation of genotypes at l^* for family i.

The pair of haplotypes for individual j in family i for the K observed loci is denoted by $H_{ij} = (H_{ij}^1, H_{ij}^2)$. We denote the possible haplotypes at K loci with numbers $1, ..., 2^K$. Haplotypes formed by observed loci and the unobserved disease locus are denoted by $H_{ij}^* = (H_{ij}^{*1}, H_{ij}^{*2})$ with 2^{K+1} possible haplotypes.

Figure 4.1: Genetic model with four observed loci and one unlinked disease locus



We denote by $\mathcal{H}(L) = \{(i,j) \in \{1,...,2^K\}^2 | i < j\}$ the possible unordered diplotypes that can be observed for loci in L (compare section 2.2.3). Analogously $\mathcal{G}(L)$ is the set of possible genotypes that can be observed at loci L. For biallelic loci, we can enumerate unordered genotypes as $g \in \{0,1,2\}$ by counting one allele in the genotype. Then, $\mathcal{G}(L) = \{0,1,2\}^K$.

In this thesis, I assume that the phase of haplotypes is unknown and diplotypes are ambiguous. As introduced in section 2.2.3, we write $h^1|h^2 = g$ if haplotypes h^1, h^2 are compatible with the genotype g without heeding parental origin, *i.e.* per locus, the alleles of the diplotype correspond to the genotype. I denote with $\mathcal{H}(g)$ the set of all unordered diplotypes compatible with g and use the notation for diplotype $h = (h^1, h^2)$:

$$\begin{aligned} & h^1|h^2 = g \\ := \; & \mathcal{H}(g) \\ := \; & \{(h^1,h^2)\,|(h^1,h^2) \in \pi_g^{-1}(g)\}. \end{aligned}$$

using definitions from (2.1), where π_g is a mapping ,,destroying" information on phase. A convenient notation for summation over $\mathcal{H}(g)$ is:

$$\sum_{h^1|h^2=g} F(h),$$

where F is some function of diplotype $h = (h^1, h^2)$ and (h^1, h^2) index the set $\mathcal{H}(g)$. Analogously, I use the notation

$$\mathbf{h_i}^{*1}|\mathbf{h_i}^{*2} = \mathbf{g_i},$$

which denotes the set of joint diplotypes at loci $L \cup L^*$, for which diplotypes are compatible with observed genotypes $\mathbf{g_i}$ after the removal of locus L^* from diplotypes $\mathbf{h_i}^{*1}|\mathbf{h_i}^{*2}$ in family i.

4.1.2. *Linkage disequilibrium.* In the general population, haplotypes are assumed to arise from a multinomial distribution with probabilities $P(H^1_{ij} = h_k) = \eta_k$ for $k = 1, ..., 2^K$ and $P(H^{*1}_{ij} = h^*_k) = \eta^*_k$ for $k = 1, ..., 2^{K+1}$, respectively. Under the assumption of Hardy-Weinberg equilibrium (HWE), $P(H_{ij} = (h_1, h_2)) = (\eta_{h_1})^2$ if $h_1 = h_2$ and $2\eta_{h_1}\eta_{h_2}$ otherwise with analogous expressions for $P(H^*_{ij})$. This assumptions is formally fulfilled in panmictic populations, *i.e.* mating is at random and each haplotype in a diplotype is independently drawn from the previous population.

The haplotype distribution can be re-parametrized in terms of the correlation structure of alleles on haplotypes. For pairs of loci the linkage disequilibrium (LD) coefficient is defined as the covariance between allele occurrence. Assume two loci with alleles A, a and B, b and let p_A and p_B denote the respective allele frequencies.

$$\delta_{AB} = Cov(I\{A\}, I\{B\}) = p_{AB} - p_A p_B, \tag{4.1}$$

when p_{AB} is the probability of haplotype AB and $I\{A\}$, $I\{B\}$ are indicator variables

We note that

$$\begin{aligned}\forall i,j: \sum_k \delta_{ik} &= \sum_k p_{ik} - \sum_k p_i p_k = p_i - p_i = 0 = \sum_k \delta_{kj} \\ \delta_{ij} &= -\sum_{i' \neq i} \delta_{i'j} \\ \delta_{ij} &= -\sum_{j' \neq j} \delta_{ij'}\end{aligned}$$

Therefore, in the bi-allelic two-locus setting a single LD parameter and the marginal frequencies characterize the joint distribution. The LD can be standardized in several ways, for example to a maximal value of 1 [58]:

$$\delta'_{AB} := \frac{\delta_{AB}}{min(p_A, p_B) - p_A p_B} \tag{4.2}$$

It is also common to consider the square of the correlation between the indicator variables on the alleles which naturally confines this parameter between 0 and 1 [22]:

$$R^2 = corr(I\{A\}, I\{B\})^2. \tag{4.3}$$

Parametrizations for more than two loci are discussed in section 5.2.1.

4.2. **Assumptions used in the likelihood framework in this thesis.**

(A1) Given the causal locus in the observed region, phenotype distributions are independent for different (related) individuals.

(A2) Mendelian inheritance holds and is independent of phenotypes.

(A3) Hardy-Weinberg equilibrium is assumed to hold for diplotypes.

(A4) There is no recombination between observed loci and the disease locus.

(A5) A genetic effect exists, i.e. $\beta > 0$ for the logistic penetrance function.

(A6) The baseline penetrance (μ) is known.

4.2.1. *Assumption (A1).* This assumption expresses the fact that the penetrance function is determined by genetic information of the observed region alone. This assumption ignores certain environmental effects and other loci influencing the phenotype and is relaxed by the model that includes a random effect.

4.2.2. *Assumption (A2).* This is a standard assumption throughout statistical genetics. (A2) implies that the selection of the gamete that is passed along is a Bernoulli process with probability $\frac{1}{2}$. A deviation from this distribution is called segregation distortion, which, while present in the human genome, only seems to have small effects [112]. This assumption is only necessary for ascertained samples, as in random samples segregation distortion is independent of phenotypes by definition.

4.2.3. *Assumption (A3).* Hardy-Weinberg equilibrium is a very common assumption since it has been shown to hold in many samples [109]. This assumption allows us to model genotype/diplotype frequencies in terms of allele/haplotype frequencies. The likelihood is continuous in deviations from HWE (as for example measured by a deviation parameter [28]), such that small violations of HWE should result in small biases in parameter estimates. From a practical perspective family based studies might have the advantage of not involving two groups as compared to the case control design, thereby arguably reducing the risk of stratification.

4.2.4. *Assumption (A4).* Recombinations are rare events on a small genetic scale. Roughly, a 1% chance of recombination equals 1Mb of DNA. If the recombination fraction is substantially larger than 0, LD patterns quickly tend to zero and it is therefore reasonable to assume $\theta \approx 0$ for the situations relevant in our setting. Larger family structures, *i.e.* not just nuclear families, would be required to include recombination into this model.

4.2.5. *Assumption (A5).* This assumption is to ensure identifiability of the likelihood presented later. Informally, if no genetic effect exists ($\beta = 0$), it is not meaningful to estimate allele frequencies of disease alleles. In a practical setting, a genetic effect should be assured, *e.g.* by comparing phenotype correlations between relatives and unrelated individuals (heritability studies, [67]). Heritabilities are established for most common diseases [31]. Also, intuitively, some genetic influence should exist for almost any trait.

4.2.6. *Assumption (A6).* Assuming the baseline penetrance parameter (μ) makes some assumption about the expected genetic effect. If the genetic effect is small, the baseline penetrance is similar to the disease prevalence $P(Y = 1)$ and can therefore be inferred from epidemiological studies.

Assumptions (A2) - (A6) are needed in all models to follow. Assumption (A1) is relaxed in the random effects models.

5. Likelihood framework

In this chapter, the likelihood used in this thesis is introduced and differences to models from chapter 3 are discussed. The likelihood is presented in terms of two parts, the penetrance model and the segregation part. The penetrance model connects causal genotypes with phenotypes and segregation models the likelihood of transmitted haplotypes. A proof of identifiability of the model without random effect is given in section 6.

5.1. **The Penetrance Model.** The probability of disease of individual ij depends on the unobserved genotype G^*_{ij} at the latent disease locus with alleles a and A. We assume that A is the disease associated allele with allele frequency $p^* = P(A)$. The following scores model the genotype for an assumed mode of inheritance: $X^\star_{ij} = 0$ for $G^\star_{ij} = \{aa\}$, $X^\star_{ij} = k_2$ for $G^\star_{ij} = \{AA\}$, and $X^\star_{ij} = k_1$ for $G^\star_{ij} = \{aA\}$, where $k_1 = 0, k_2 = 1$ for a recessive mode of inheritance, $k_1 = k_2 = 1$ for a dominant mode of inheritance, and $k_1 = 1, k_2 = 2$ for an additive mode of inheritance.

The penetrance function is given by

$$\text{logit}(p_{ij}) = \text{logit}\{P(Y_{ij} = 1|X^\star_{ij}, a_i)\} = \mu + \sigma_a a_i + \beta X^\star_{ij} \tag{5.1}$$

Here, μ is the common intercept, that we assume is known from external data. The family specific random intercept a_i allows for different baseline penetrances for different families, and follows a normal distribution with mean 0 and variance 1. σ_a is a parameter scaling the variance of the random effect. For identifiability purposes, we assume that $\beta \neq 0$, *i.e.* the disease truly has a genetic component. Implications of this assumption are discussed later. The above penetrance function can be extended to include other measured covariates Z_{ij} using $\text{logit}\{P(Y_{ij} = 1|X^\star_{ij}, a_i, Z_{ij})\} = \mu + \sigma_a a_i + \beta X^\star_{ij} + \gamma Z_{ij}$.

Under the logistic model (5.1), the marginal probability of the response in the ith family requires integration over the random effects distribution, an operation that cannot be carried out in closed form, and is written as

$$P(\mathbf{Y_i}|\mathbf{X_i^*}) = P(Y_{i1},\dots,Y_{in_i}|X_{i1}^\star,\dots,X_{in_i}^\star) = \int \prod_{j=1}^{n_i} p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} dF(a_i), \tag{5.2}$$

where $q_{ij} = 1 - p_{ij}$. We also note that G^* is the only genotype influencing disease risk in the region, thus leading to conditional independence of phenotypes and the observed genotypes:

$$P(\mathbf{Y_i}|\mathbf{G_i},\mathbf{G_i^*}) \;=\; P(\mathbf{Y_i}|\mathbf{G_i^*}) \;=\; P(\mathbf{Y_i}|\mathbf{X_i^*}) \;=\; P(Y_{i1},\dots,Y_{in_i}|X_{i1}^\star,\dots,X_{in_i}^\star) \tag{5.3}$$

When $\sigma_a^2 = 0$, model (5.1) reduces to standard logistic regression and $P(\mathbf{Y_i}|\mathbf{X_i^*}) = \prod_{j=1}^{n_i} P(Y_{ij}|X_{ij}^*)$.

5.2. **Likelihood for a candidate region.** First, we assume that families in our study are a random sample of families in the population. While our model can be applied to various family structures, for ease of exposition we assume each family consists of two parents and $n_i - 2$ offspring.

Using (5.3), Mendelian inheritance, the assumption of no recombinations in the observed region, and the fact that the genetic contributions of offspring are independent, conditional on the parental pair of haplotypes, the likelihood for N families is obtained by summation over the true unobserved disease genotype. In (5.6), we use the recipe from section 2.2.3 to express the likelihood in terms of haplotypes. (5.7) seperates terms into the penetrance model and the likelihood for diplotypes and conditions on parental haplotypes. In (5.8) the indendence of offspring diplotypes given parental diplotypes is used. Likelihood terms for offspring diplotypes only depend on Mendelian inheritance:

$$L(\theta) = P(\mathbf{Y}, \mathbf{G}, \theta) = \prod_{i=1}^{N} P(\mathbf{Y_i}, \mathbf{G_i}) = \prod_{i=1}^{N} \sum_{\mathbf{g_i^\star}} P(\mathbf{Y_i}, \mathbf{G_i}, \mathbf{G_i^*} = \mathbf{g_i^*}) \tag{5.4}$$

$$= \prod_{i=1}^{N} \sum_{\mathbf{h_i^*} \in \mathbf{h_i^{*1}} | \mathbf{h_i^{*2}} = \mathbf{g_i}} P(\mathbf{Y_i}, \mathbf{H_i^*} = \mathbf{h_i^*}) \tag{5.5}$$

$$= \prod_{i=1}^{N} \sum_{\mathbf{h_i^{*1}} | \mathbf{h_i^{*2}} = \mathbf{g_i}} P(\mathbf{Y_i} | \mathbf{G_i^*} = \mathbf{g_i^*}) P(H_{i3}^* = h_{i3}^*, \ldots, H_{in_i}^* = h_{in_i}^* | h_{i1}^*, h_{i2}^*) \tag{5.6}$$

$$\times P(H_{i1}^* = h_{i1}^*) P(H_{i2}^* = h_{i2}^*) \tag{5.7}$$

$$= \prod_{i=1}^{N} \sum_{\mathbf{h_i^{*1}} | \mathbf{h_i^{*2}} = \mathbf{g_i}} P(\mathbf{Y_i} | \mathbf{g_i^*}) \prod_{j=3}^{n_i} P(H_{ij}^* = h_{ij}^* | h_{i1}^*, h_{i2}^*) \prod_{j=1}^{2} P(H_{ij}^* = h_{ij}^*). \tag{5.8}$$

The set of haplotypes compatible with the observed genotypes $\mathbf{g_i}$ is denoted by $\mathbf{h_i^{*1}} | \mathbf{h_i^{*2}} = \mathbf{g_i}$ and $\mathbf{g_i^\star}$ is the set of possible genotype combinations at locus l^* for family i. Note, that $\mathbf{h_i^{*1}} | \mathbf{h_i^{*2}}$ includes the disease locus, whereas $\mathbf{g_i}$ is restricted to observed loci L (see section 4.1.1).

The penetrance function $P(\mathbf{Y_i} | \mathbf{G_i^*})$ depends on parameters β and σ_a as defined in (5.1). Given the parental pair of haplotypes, the genetic contributions of offspring $P(H_{ij}^* | h_{i1}^*, h_{i2}^*)$ follow from Mendelian inheritance and thus do not need to be parametrized. There are several parametrizations that can be used for the parental genetic information $P(\mathbf{H}^*)$, all depending on $2^{K+1} - 1$ parameters, where K denotes the number of observed loci.

It remains to parametrize the parental haplotype distribution. In the following, four parametrizations of the parental haplotype distribution are presented that are later used for different purposes.

5.2.1. *Parameterization 1.* Haplotypes including loci L and $L^\star$ are considered jointly and are parametrized by a multinomial $\mathcal{M} = M(1, \eta_1^*, ..., \eta_{2^{K+1}}^*)$. Parental haplotypes $H_{11}^{*1}, H_{11}^{*2}, H_{12}^{*1}, H_{12}^{*2}, ..., H_{N1}^{*1}, H_{N1}^{*2}, H_{N2}^{*1}, H_{N2}^{*2}$ are *iid* $H_{ij}^{*k} \sim \mathcal{M}, i = 1, ..., N; j = 1, 2; k = 1, 2$.

The parameters and the corresponding terms of the likelihood are listed in the following table.

- β, σ_a : Penetrance function $P(Y_i|g_i^\star)$
- $\eta_1^\star, ..., \eta_{2^{K+1}}^\star, \sum \eta_i^\star = 1$: Parental haplotypes

The vector of free parameters is then given by $\theta_1 = (\eta_1^\star, ..., \eta_{2^{K+1}-1}^\star, \beta, \sigma_a)$ with $2^{K+1}+1$ entries. This parametrization is used in the Bayesian formulation of the problem in section 7.

5.2.2. *Parameterization 2.* This parameterization separates observed and the latent locus. We write haplotype H_{ij}^{*k} as $H_{ij}^{*k} = (H_{ij}^k, G_{ij}^{*k})$, and

$$P(H_{ij}^{*k} = h_{ij}^{*k}) = P((H_{ij}^k, G_{ij}^{*k}) = (h_{ij}^k, g_{ij}^{*k})) = P(G_{ij}^{*k} = g_{ij}^{*k}|h_{ij}^k)P(H_{ij}^k = h_{ij}^k).$$

I parametrize haplotypes at observed loci H_{ij}^k with parameters of a multinomial $\eta_1, ..., \eta_{2^K}$ and conditional probabilities $P(G_{ij}^{*k} = g_{ij}^{*k}|h_{ij}^k)$ with $\xi_{h_{ij}^k}$ resulting in $\xi_1, ..., \xi_{2^K}$. The parameters and the corresponding terms of the likelihood are listed in the following table.

- β, σ_a : Penetrance function $P(Y_i|g_i^\star)$
- $\eta_1, ..., \eta_{2^K}$: Haplotypes at observed loci
- $\xi_1, ..., \xi_{2^K}$: Conditional allele frequencies $P(G^\star|h_i)$

Because of the constraint $\sum \eta_i = 1$, we have free parameters $\theta_2 = (\eta_1, ..., \eta_{2^K-1}, \xi_1, ..., \xi_{2^K}, \beta, \sigma_a)$. This parametrization is used in the proof of identifiability (section 6).

5.2.3. *Parameterization 3.* One of the goals of this work is to estimate linkage disequilibrium (LD) between the disease allele and haplotypes at the observed loci. For two biallelic loci with minor alleles A and B, with allele frequencies p_A and p_B and haplotype frequency p_{AB} of AB, δ is defined as $\delta = p_{AB} - p_A p_B$ [96]. Recall, that p^* denotes the allele frequencies of allele A at the latent disease locus with genotype G^*. The joint distribution of H^* is characterized by LD between haplotypes H corresponding to the observed genotypes and the disease allele A, *i.e.* $\delta_h = p_{hA} - p_h p_A = Cov(I\{g^* = A\}, I\{H = h\}))$ for haplotype h. The δs satisfy the constraint $\sum_{i=1}^{(2^K)} \delta_i = 0$. One parametrization of the likelihood (5.4) is therefore given by $\theta_3 = (\eta_1, ..., \eta_{2^K-1}, \delta_1, ..., \delta_{2^K-1}, p^*, \beta, \sigma_a)$. The parameters and the corresponding terms of the likelihood are listed in the following table.

• β	: Penetrance function $P(Y_i \vert g_i^\star)$
• $\eta_1, ..., \eta_{2^K}$	: Haplotype frequencies at observed loci
• $\delta_1, ..., \delta_{2^K}$	: pairwise covariance between haplotypes at L and disease allele
• $p^\star$	: Allele frequency of disease allele at $L^\star$.

This parametrization allows for a straightforward testing of the null hypothesis of $H_0 : \delta_1, ..., \delta_{2^K} = 0$ which corresponds to no genetic association.

5.2.4. *Parameterization 4.* As an alternative to parametrization 3 using δ, I consider the correlation [96] between haplotype h and the disease allele,

$$R_i = \frac{\delta_i}{\sqrt{p^*(1-p^*)}\sqrt{\eta_i(1-\eta_i)}}.$$

and parametrize (5.8) using $\theta_4 = (\eta_1, ..., \eta_{2^K-1}, R_1, ..., R_{2^K-1}, p^*, \beta, \sigma_a)$. Again, the parameters and the corresponding terms of the likelihood are listed in the following table.

• β	: Penetrance function
• $\eta_1, ..., \eta_{2^K}$	: Haplotype frequencies at observed loci
• $R_1, ..., R_{2^K}$	: pairwise correlation between haplotypes at L and disease allele
• $p^\star$	: Allele frequency of disease allele at $L^\star$

Analogously to parametrization 3, this parametrization allows for straightforward testing of the null hypothesis of $H_0 : \delta_1, ..., \delta_{2^K} = 0$ which corresponds to no genetic association. The interpretation of $R = (R_1, ..., R_{2^K})$ can be more intuitive than $\delta = (\delta_1, ..., \delta_{2^K})$ and is therefore preferred in the simulations.

5.2.5. *Proof of equivalence.* For convenient notation we omit indices for haplotypes in this section, *i.e.* we write H for H_{11}^1 and H^* for $H_{11}^{*1} = (H_{11}^1, G_{11}^{*1}) = (H, G^*)$ making use of the fact that parental haplotypes are *iid* as $\mathcal{M}$.

■ **Lemma 5.1.** *Parametrization 1, 2, 3 and 4 are equivalent, i.e. there exists a differentiable bijection between the parameter sets.*

Proof. The goal is to show that the various parametrization of $H^\star$ are equivalent. Let K be the number of observed loci, then there are $M = 2^{K+1}$ possible haplotypes at loci $L \cup L^\star$.

"1 ↔ 2": Define

$$\begin{aligned}
f_{12} : \Theta_1 \to \Theta_2 \quad &: \quad (\eta_1^\star, ..., \eta_M^\star) \to (\eta_1, ..., \eta_{M/2}, \xi_1, ...\xi_{M/2}) \\
\Theta_1 \quad &= \quad (\eta_1^\star, ..., \eta_M^\star) \in (0,1)^M, \sum_{i=1}^{M} \eta_i^\star = 1 \\
\Theta_2 \quad &= \quad (\eta_1, ..., \eta_{M/2}, \xi_1, ...\xi_{M/2}) \in (0,1)^M, \sum_{i=1}^{M/2} \eta_i = 1 \\
\eta_i \quad &:= \quad \eta_i^\star + \eta_{i+M/2}^\star \\
\xi_i \quad &:= \quad \eta_i^\star / \eta_i
\end{aligned}$$

The definition of f_{12} is based on the equation: $\eta_h = P(H = h) = \sum_{G^\star = g^\star} P(H^\star = h^\star) = \{P(H = h, G^\star = a) + P(H = h, G^\star = A)\} = \eta_{ha}^* + \eta_{hA}^*$, with "$ha$" corresponding to some i and "hA" corresponding to $i + M/2$. We note that η_i and ξ_i only depend on $\eta_i^\star, \eta_{i+M/2}^\star$. Without loss of generality (WLOG) we only consider η_1 and ξ_1 which depend on $\eta_1^\star, \eta_{1+M/2}^\star =: \eta_1^{\star'}$. The inverse of f_{12} is readily given by:

$$f_{12}^{-1}(\eta_1, \xi_1) \quad = \quad (\eta_1 \xi_1, \eta_1(1 - \xi_1)) = (\eta_1^\star, \eta_1^{\star'})$$

This shows that f_{12} is injective. To show that f_{12} is surjective let us assume that for full parameter vectors $\eta^\star = f_{12}^{-1}(\eta, \xi)$. Then $\sum_{i=1}^{M} \eta_i^\star = \sum_{i=1}^{M/2} \eta_i(\xi_i + (1 - \xi_i)) = \sum_{i=1}^{M/2} \eta_i = 1$, *i.e.* $\eta^* \in \Theta_1$ which demonstrates that f is surjective. Furthermore f_{12} is a rational function in each component, which implies $f_{12} \in C^\infty$.

"$1 \leftrightarrow 3$": Define

$$\begin{aligned}
f_{13} : \Theta_1 \to \Theta_3 \quad &: \quad (\eta^\star_1, ..., \eta^\star_M) \to (\eta_1, ..., \eta_{M/2}, \delta_1, ...\delta_{M/2}, p^\star) \\
\Theta_1 \quad &= \quad (\eta^\star_1, ..., \eta^\star_M) \in (0,1)^M, \sum_{i=1}^M \eta^\star_i = 1 \\
\Theta_3 \quad &= \quad (\eta_1, ..., \eta_{M/2}, \delta_1, ...\delta_{M/2}, p^\star) \in (0,1)^{M/2} \times (b^1_1, b^2_1) \times ... \times (b^1_{M/2}, b^2_{M/2}) \times (0,1), \\
& \quad \quad \sum_{i=1}^{M/2} \eta_i = 1, \sum_{i=1}^{M/2} \delta_i = 0, \\
& \quad \quad b^1_i = max(\eta_i + p^\star - 1, 0) - \eta_i p^\star, b^2_i = min(\eta_i, p^\star) - \eta_i p^\star \\
\eta_j \quad &:= \quad \eta^\star_j + \eta^\star_{j+M/2}, j = 1, ..., M/2 \\
\delta_j \quad &:= \quad \eta^\star_{M/2+j} - \eta_j p^\star, j = 1, ..., M/2 \\
p^\star \quad &:= \quad \sum_{i=M/2+1}^{M} \eta^\star_i
\end{aligned}$$

For haplotype h, δ_h is therefore $\delta_h = Cov(I\{H = h\}, I\{G^* = A\})$. As above WLOG, we consider frequencies $\eta^\star_1, \eta^\star_{1+M/2}$ and use the above notation. The inverse of f_{13} is given by:

$$f^{-1}_{13}(\eta_1, \delta_1, p^\star) \quad = \quad (\eta_1 - (\delta_1 + \eta_1 p^\star), \delta_1 + \eta_1 p^\star) = (\eta^\star_1, \eta^{\star'}_1) \in (0,1)^2.$$

For given $(\eta, \delta, p^\star)$ we now demonstrate that $f^{-1}_{13}(\eta, \delta, p^\star) \in \Theta_1$. $\sum \eta^\star_i = \sum_{i=1}^{M/2}(\eta_i - (\delta_i + \eta_i p^\star)) + \sum_{i=M/2+1}^{M}(\delta_i + \eta_i p^\star) = \sum_{i=1}^{M} \eta_i = 1$. On account of being rational $f_{13} \in C^\infty$.

"$3 \leftrightarrow 4$": Equivalence of parametrization 3 and 4 is given by the conversion of covariance to correlation for fixed marginals. □

5.3. **Extensions.**

5.3.1. *Multiple unobserved loci.* In a realistic scenario, multiple genes have to be assumed to influence a given phenotype. If a region is observed, it is in linkage equilibrium with most of the genome. These effects can be modeled by means of random effects. I here characterize two possibilities, contrasting the model presented above with a more complex model.

5.3.2. *Random effects model for unobserved loci.* To account for additional loci in linkage equilibrium affecting the phenotype, a random effect can be used [77]. For a logistic penetrance function, the model is:

$$\text{logit}(P(Y_{ij} = 1|G^{\star}_{ij} = g^{\star}_{ij}, a_{ij})) = \mu + \beta x^{\star}_{ij} + \sigma a_{ij} \tag{5.9}$$

Here, μ is the known intercept, a_{ij} is the random effect of individual j in the ith family and $x^{\star}_{ij}$ is the score of $g^{\star}_{ij}$, according to the assumed mode of inheritance. The a_{ij} are normally distributed with mean 0 and a covariance matrix that can be parametrized in terms of relationship degrees in the family. The covariance of two individuals is given by the probability that a given locus is shared IBD (or by how much genetic material two individuals have in common on average). This probability is called the kinship coefficient (as in segregation analysis, section 3.4). For first degree relatives as sibs and parent-offspring pairs the kinship coefficient is $\frac{1}{2}$. For second degree relatives it is $\frac{1}{4}$. Therefore the covariance matrix Σ for a family consisting of parents and offspring of $a_i = (a_{i1}, ..., a_{in_i})$ is given by:

$$\Sigma = \Big(\text{Cov}(a_{ij}, a_{ik})\Big)^{(n_i,n_i)}_{(j,k)=(1,1)} = \begin{pmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} & 1 \end{pmatrix}$$

The parameter σ scales the effect of the random effect such that $Cov(\sigma a_i) = \sigma^2\Sigma$. A downside of this model is its computational complexity. There is no symbolic form of the integral

$$\int L(\theta; Y, G)dF(a)$$

that needs to be computed to evaluate the likelihood. Here L is the likelihood from (5.4) with the penetrance function replaced as specified above, $a = (a_1, ..., a_N)$ and θ is the parameter vector including σ. Evaluating the multidimensional integral by numeric Monte Carlo integration becomes infeasible for family sizes of about 5. I therefore consider the simplified random effects model as introduced in section 5.1:

$$\text{logit}(P(Y_{ij} = 1 | G^{\star}_{ij} = g^{\star}_{ij}, a_i)) = \mu + \beta x^{\star}_{ij} + \sigma_a a_i, \tag{5.10}$$

with $a_i \sim \text{iid}\, N(0, 1)$. Thereby a single random effect per family is introduced. This random effect requires to evaluate a one-dimensional integral for which numerical quadrature methods are applicable. The implementation of the random effect turns out to be about 100 times slower than the model without random effect, which is much faster than an expected factor of $10^5 - 10^7$ for a Monte Carlo integration for a family size of, say 5, for the multi-dimensional random effect.

With respect to interpretation, both random effects models have their advantages. While model (5.9) exactly models dependencies according to segregation, model (5.10) might be able to better capture shared environmental effects in situations where exposition is homogeneous within families (*e.g.* diet).

5.3.3. *Covariates.* Covariates can be included into the model by modifying the penetrance function. For covariates Z_{ij} of individual j of family i, a term $\gamma' Z_{ij}$ can be introduced:

$$\text{logit}(P(Y_{ij} = 1 | G^{\star}_{ij} = g^{\star}_{ij}, a_i)) = \mu + \beta x^{\star}_{ij} + \sigma_a a_i + \gamma' Z_{ij},$$

where γ would be the corresponding parameter vector.

5.3.4. *Ascertainment.* To enrich the sample with affected individuals, it is common to ascertain families on having at least one affected offspring. The likelihood is now conditional on the ascertainment event A:

$$L(Y,G|A) = \frac{L(Y,G,A)}{L(A)},$$

for the log-likelihood we have, given observations fulfill condition A:

$$\log L(Y,G|A) = l(Y,G|A) = l(Y,G) - l(A)$$

If A is the event that at least one offspring is affected, i.e. $A = \{\sum_{i=1}^{k} Y_i \geq 1\}$, the ascertainment correction $L(A)$ is given by a marginalizing over parental genotypes:

$$\begin{aligned} L(A) &= P(\sum_{i=1}^{k} Y_i \geq 1) = 1 - P(\sum_{i=1}^{k} Y_i = 0) \\ &= 1 - \sum_{G_1=g_1,G_2=g_2} P(g_1,g_2) \prod_{i=1}^{k} \sum_{g_i} P(Y_i = 0, G_i = g_i | g_1, g_2). \end{aligned}$$

5.4. **Parameter estimation.** Parameter estimates can be obtained by maximum likelihood (ML) estimation. The structure of the likelihood does not allow to infer a closed form solution for the estimates, and therefore numerical optimization is used. Constraints on the parameters are handled by remapping parameters of the likelihood[6]. Details of the implementation are given in appendix C.1 (page 111).

5.4.1. *Confidence intervals.* Confidence intervals can be constructed using asymptotic normal theory. Asymptotically, the ML-estimators are normally distributed ($\sqrt(n)(\hat{\theta} - \theta) \to N(0, I^{-1})$). An α-level confidence interval for parameter θ can be constructed as follows:

$$CI(\hat{\theta}_i, \alpha) = \left(\hat{\theta}_i - z_{\alpha/2}\sqrt{(\hat{I}^{-1})_{ii}}, \hat{\theta}_i + z_{\alpha/2}\sqrt{(\hat{I}^{-1})_{ii}}\right),$$

[6] The R function *optim* is used with the Nelder-Mead algorithm. It is sufficient to return a value of $-\infty$ outside the valid parameter space to ensure valid estimates. This approach was finally used, since it has the same effect as computing the likelihood on the border of the parameter space and adding a penalty term according to the distance to the border for given parameter values.

where α is the level of the confidence interval, $\hat{I}$ is the estimated Fisher information matrix and $z_{\alpha/2}$ is the $\alpha/2$ quantile of the normal distribution. As this result only holds asymptotically, the speed of convergence needs to be checked for finite samples. This is done later in section 8.6. An alternative to confidence intervals constructed from asymptotic normality are bootstrap confidence intervals [18] which are employed in the data analysis (section 9).

5.5. **Statistical testing.** A key assumption of the model is that a genetic effect exists, *i.e.* $\beta \neq 0$ in the penetrance function. The appropriate null hypothesis is therefore that there is no disease locus linked to observed loci in a region, *i.e.* $H_0 : \delta_1 = ... = \delta_M = 0$. This test scenario can be evaluated by means of a likelihood ratio test:

$$T = 2\left(\sup_{\theta \in \Theta_1} l(\theta|Y,G) - \sup_{\theta \in \Theta_0} l(\theta|Y,G)\right),$$

when T is asymptotically χ^2 distributed with M degrees of freedom. $M = 2^K - 1$ is the number of parameters δ_i and is defined by the number k of observed loci.

In the case of one observed and one unobserved disease locus, the null hypothesis of no association can be expressed by $\delta = 0$ as outlined in section 2.2.6.

Again, the χ^2_M distribution holds only asymptotically. α-level simulations are conducted in section 8.6 to assess convergence in finite samples.

6. Properties of the likelihood

In this section, the identifiability of the likelihood (5.4; p. 40) is shown. As the model has several parameters connected to unobserved variables, identifiability is not clear outright. In fact, it turns out to be non-trivial to show identifiability which is proven for the simplified model (5.4) with $\sigma_a = 0$. The problem is tackled by representing the likelihood in a factorial way for which a marginal and conditional part can be considered separately. This chapter is organized as follows: first the decomposition is presented in a general framework, then this decomposition is applied to our likelihood and identifiability is shown by applying the general concepts. Finally, consistency of MLEs is shown and the case of $\sigma_a \neq 0$ is discussed.

6.1. **Identifiability of conditionally decomposable likelihoods.** The motivation to consider decompositions is given by the fact that likelihood (5.4) can also be written as $P(Y, G = g) = P(Y|G = g)P(G = g)$. Let $X : \Omega \to R \supset T, Y : \Omega \to S$ be random variables. In this section we consider likelihoods of the form

$$P_\theta(X, Y) = P_\alpha(X)P_\beta(Y|X = x), \tag{6.1}$$

for which $\theta = (\alpha, \beta)$. Let us now assume that $P_\alpha(X)$ is identifiable and $P_\beta(Y|X = x)$ is identifiable for all $x \in T$. The identifiability of $P_\alpha(X)$ and $P_\beta(Y|X = x)$ implies the identifiability of $P_\theta(X, Y)$ as we will show that $P_\theta(X = x, Y = y) = P_{\theta'}(X = x, Y = y)\,\forall x, y \Rightarrow \theta = \theta'$. Let us assume that

$$\begin{aligned} & \forall x, y : \; P_\theta(X = x, Y = y) = P_{\theta'}(X = x, Y = y) \\ \Leftrightarrow \quad & \forall x, y : \; P_\beta(Y = y|X = x)P_\alpha(X = x) = P_{\beta'}(Y = y|X = x)P_{\alpha'}(X = x). \end{aligned}$$

From this follows:

$$\begin{aligned}\forall x: \qquad P_\alpha(X=x) &= \sum_x \{P_\beta(Y=y|X=x)P_\alpha(X=x)\} \\ &= \sum_x \{P_{\beta'}(Y=y|X=x)P_{\alpha'}(X=x)\} = P_{\alpha'}(X=x),\end{aligned}$$

such that $\alpha = \alpha'$ by the identifiability of $P_\alpha(X)$. With $\alpha = \alpha'$, we have:

$$\begin{aligned}&\forall x, y: \ P_\beta(Y=y|X=x)P_\alpha(X=x) = P_{\beta'}(Y=y|X=x)P_\alpha(X=x) \\ \Leftrightarrow \ &\forall x, y: \ P_\beta(Y=y|X=x) = P_{\beta'}(Y=y|X=x),\end{aligned}$$

which implies $\beta = \beta'$ by the assumption of identifiability of $P_\beta(Y=y|X=x)$ for all $x \in T$. As a result $\theta = \theta'$, thereby concluding the argument. In the next section likelihood (5.8) is rearranged such as to be of the form $P_\beta(Y=y|X=x)P_\alpha(X=x)$ and identifiability of the components is shown.

6.2. **Identifiability of likelihood** L. We assume parametrization 2 (section 5.2.2), *i.e.* $\theta = \theta_2 = (\eta_1, ..., \eta_{2^K-1}, \xi_1, ..., \xi_{2^K}, \beta, 0)$. Identifiability is shown for the likelihood for a randomly sampled family (5.8). Without loss of generality we assume a single sampled family, *i.e.* $N = 1$ with k offspring, *i.e.* $n_1 = k$. We start by rewriting the likelihood as follows:

$$\begin{aligned}L(\theta) = P(\mathbf{Y}, \mathbf{G}, \theta) = P(\mathbf{Y_1}, \mathbf{G_1}) &= \sum_{\mathbf{g_1^\star}} P(\mathbf{Y_1}, \mathbf{G_1}, \mathbf{G_1^*} = \mathbf{g_1^*}) \\ &= P(\mathbf{G_1}) \sum_{\mathbf{g_1^\star}} P(\mathbf{Y_1}, \mathbf{G_1^*} = \mathbf{g_1^*} | \mathbf{G_1} = \mathbf{g_1}) \\ &= \underbrace{\sum_{\mathbf{h_1^1}|\mathbf{h_1^2}=\mathbf{g_1}} P(\mathbf{H}_1 = \mathbf{h}_1)}_{L_1} \underbrace{\sum_{\mathbf{g_1^\star}} P(\mathbf{Y_1}|\mathbf{G_1^*} = \mathbf{g_1^*}) P(\mathbf{G_1^*} = \mathbf{g_1^*}|\mathbf{G_1} = \mathbf{g_1})}_{L_2}. \qquad (6.2)\end{aligned}$$

Note, that the probability $P(\mathbf{G_1^*} = \mathbf{g_1^*}|\mathbf{G_1} = \mathbf{g_1})$ depends on all diplotypes compatible with $\mathbf{g_1}$, which is more complex a formulation than in (5.4). In the following WLOG

Table 6.1: Enumeration of set M, the genotypes/diplotypes/haplotypes for homozygous genotypes. We denote alleles at locus $\mathcal{A}$ and $\mathcal{B}$ with a, A and b, B, respectively. The resulting unambiguous haplotype and diplotype are given in respective columns. Column *Outcome* iterates the possible outcomes, and column *Parameter* lists the parameter the realization depends on.

Genotype $\mathcal{A}$	Genotype $\mathcal{B}$	Haplotype	Diplotype	Outcome	Parameter
(a,a)		a	(a,a)	1	η_1
(A,A)		A	(A,A)	2	η_2
(a,a)	(b,b)	ab	(ab,ab)	1	η_1
(a,a)	(B,B)	aB	(aB,aB)	2	η_2
(A,A)	(b,b)	Ab	(Ab,Ab)	3	η_3
(A,A)	(B,B)	AB	(AB,AB)	4	η_4

we restrict our considerations to outcomes for $\mathbf{G_1}$ which are confined to homozygous genotypes at all loci, which additionally are identical for all family members. Table 6.1 iterates possible outcomes for one and two loci. We denote this set M and therefore assume $G_{1j} \in M \,\forall j$. Under this constraint, genotypes correspond unambiguously to haplotypes and diplotypes in a one-by-one correspondence, thus allowing to iterate outcomes in a straightforward fashion as apparent in the *Outcome* column. We call L_1 the segregation part and L_2 the phenotype part of L and deal with these components according to section 6.1. Observe, in this vein, that L_1 only dependes on $\alpha = (\eta_1, ..., \eta_{2^K-1})$ and L_2 only depends on $\gamma = (\xi_1, ..., \xi_{2^K}, \beta)$.

6.2.1. *Identifiability of the segregation part of the likelihood.* In order to establish identifiability of $L(\theta_2)$, we have to show identifiability of the likelihood

$$L_1(\alpha; \mathbf{G}_1) = \sum_{\mathbf{h_1^1}|\mathbf{h_1^2}=\mathbf{g_1}} P(\mathbf{H}_1 = \mathbf{h}_1).$$

For outcomes $\mathbf{g} = (g, ..., g); g \in M$ we get:

$$L_1(\alpha; \mathbf{G}_1 = \mathbf{g}) = \eta_g^4, \tag{6.3}$$

as transmissions to offspring have probability 1 and L_1 therefore reduces to parental probabilities. This identifiability problem can be reduced to that of a single individual, as a single individual can be considered to be the marginal distribution of the full family (in likelihood L_1)[7]. The direct connection of observation with parameters in (6.3) proves identifiability of L_1, as clearly $\eta_g \neq \eta'_g \Rightarrow L_1(\eta_g; \mathbf{G}_1 = \mathbf{g}) \neq L_1(\eta'_g; \mathbf{G}_1 = \mathbf{g})$. Algorithmically, the likelihood of genotypes of unrelated individuals parametrized by haplotype frequencies has been extensively studied. Excoffier and Slatkin (1995) [21] established an EM-Algorithm to estimate haplotype frequencies of such a likelihood, however identifiability/consistency of ML-estimates is not shown. In a recent paper [59] identifiability of haplotype distributions is implicitly dealt with for a haplotype based association test, this model differing from likelihood L by assuming all loci in the model are observed.

6.2.2. *Identifiability of the phenotype part.* The following theorem constitutes the main difficulty in showing the identifiability of the likelihood. The proof is completed by a series of lemmata. The theorem is stated first and then some motivation for the lemmata is given.

■ **Theorem 6.1.** *The distribution defined by the likelihood $L_2(\gamma; Y|G)$ in 5.2 is identifiable, given the penetrance function is monotone in the penetrance parameter β for all genotypes and strictly monotone for at least one genotype. Additionally the penetrance function is assumed to be monotonic in allele dose*[8]*, i.e.* $(1 \geq f(\beta, 2) \geq f(\beta, 1) \geq f(\beta, 0) \geq 0)$ *and* $(f(\beta, 2) > f(\beta, 1) \text{ or } f(\beta, 1) > f(\beta, 0))$.

Without loss of generality, we can again consider a single diplotype d, which is present in all family members. This simplification is also intuitively motivated by the fact that observed loci L and $L^\star$ might be independent, *i.e.* $\delta_i \equiv 0 \forall i$ in parametrization 3, and therefore observed loci would not contain any information on L^*. By this choice L_2 would depend on $\gamma = (\xi_d, \beta)$. To stress the arbitrariness of d we denote ξ_d with p^* as it reflects the frequency of the predisposing allele at L^* in this family. I complete the theorem by

[7] If a marginal likelihood is identifiable, the full likelihood is identifiable as well, since otherwise, for some $\theta \neq \theta'$ the full likelihood would be constant for all possible outcomes which would contradict the identifiability of the marginal for $\theta \neq \theta'$.

[8] Allele dose denotes the number of one designated allele in a genotype which can assume values $0, 1, 2$. For example, allele dose of allele 1 in genotypes $(1, 1), (1, 2), (2, 2)$ would be $2, 1, 0$, respectively.

showing a series of lemmata and numeric calculations. The main idea is, that the likelihood surface has distinct monotonicity properties for certain phenotype observations; *i.e.* all offspring are either affected or unaffected.

- Lemma 6.2 shows a monotonicity property of certain polynomials where the likelihood at hand can be shown to have this form, when β is fixed, thereby showing monotonicity in p^*. Monotonicity in β is given by assumption.
- Next, remark 6.3 gives a condition for the gradient of a two-dimensional function f which is strictly monotonic in both variables. If for some vector v the scalar product of the gradient is positive we can find a neighborhood in which f increases in direction of v. This property is later used to compare the likelihood surfaces for the observations of only affected and only unaffected offspring, respectively.
- Lemma 6.4 then shows that for two two-dimensional functions f and g which are both differentiable and strictly monotonic in both variables, if we take the levelset of say f in $f(z)$, *i.e.* $M = f^{-1}(f(z))$, and the sign of the gradients of the two functions is identical in the whole levelset, then $g(t) \neq g(t')$ for $t \neq t'$, where both t and t' are in the levelset of f. This condition can be used to show identifiability if two likelihood surfaces for different outcomes fulfill the assumptions of the lemma.
- The next lemma 6.5 now applies lemma 6.2 to our likelihood and establishes monotonicity in both p^* and β.
- Corollary 6.6 now states that it suffices to check the assumptions of lemma 6.4 as applied to the likelihood to show identifiability. This turns out to be feasible only by numerical analysis.

■ **Lemma 6.2.** *Let* $f_n(p) = \sum_{i=0}^{n} c_i \binom{n}{i} p^i (1-p)^{n-i}$*. Then* $f_n(p)$ *is monotonically increasing (decreasing) if* $0 \leq c_0 \leq ... \leq c_n \leq 1$ *(*$1 \geq c_0 \geq ... \geq c_n \geq 0$*).* $f_n(p)$ *is strictly monotonically increasing (decreasing) for* $p \in [0,1]$ *if and only if* $\exists 0 \leq i < j \leq n : c_i < c_j$ *(*$\exists 0 \leq i < j \leq n : c_i > c_j$*).*

Proof. „$\nearrow$": We show that $f_n'(p) \geq 0$.

$$\begin{aligned} f_n'(p) &= \sum_{i=0}^{n} c_i \binom{n}{i} \left(ip^{i-1}(1-p)^{n-i} - p^i(n-i)(1-p)^{n-i-1}\right) \\ &= \sum_{i=0}^{n-1} d_i p^i (1-p)^{n-1-i}. \end{aligned}$$

Obviously, d_i is given as follows:

$$\begin{aligned} d_i &= \left((i+1)c_{i+1}\binom{n}{i+1} - c_i\binom{n}{i}(n-i)\right) \\ &= \frac{(i+1)c_{i+1}n!}{(i+1)!(n-i-1)!} - \frac{(n-1)c_i n!}{i!(n-i)!} \\ &= \frac{n!(n-i)c_{i+1} - n!(n-i)c_i}{i!(n-i)!} = \frac{n!}{i!(n-i)!}(c_{i+1} - c_i) \geq 0. \end{aligned}$$

It follows, that $f_n'(p) \geq 0$. If $\exists 0 \leq i < j \leq n : c_i < c_j$ it follows that some $d_i > 0$ and thus $f_n'(p) > 0 \forall p \in (0,1)$, showing strict monotonicity.

„$\searrow$": This case follows analogously. □

■ **Remark 6.3.** *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be differentiable and strictly monotonically increasing in both x with y fixed and y with x fixed. If the gradient in $z = (x, y)$ is given by $\nabla f|_z$, $v \in \mathbb{R}^2$ then:*

$$\nabla f|_z \cdot v > 0 \Rightarrow \exists \epsilon > 0 : \forall 0 < \epsilon' < \epsilon : f(z + \epsilon' v) > f(z)$$

Proof. By the differentiability of f the first order linear approximation of f in $z = (z_x, z_y)$ is given by

$$f(w) = f(z) + \nabla f|_z \cdot (w - z) + o\left(\frac{1}{\parallel w - z \parallel}\right).$$

With $v = (w - z)$ this Taylor expansion shows the remark. □

■ **Lemma 6.4.** *Let $g(x,y), f(x,y) : \mathbb{R}^2 \to \mathbb{R}$ be differentiable and strictly monotonically increasing in both x with y fixed and y with x fixed. Let $M = M(z)$ be the levelset of f with value $f(z)$: $M(z) = f^{-1}(f(z)) = \{v \in \mathbb{R}^2 | f(v) = f(z)\}$.*[9] *If the scalar product of*

[9] Here, sgn is the sign function and $v^\perp$ is the vector perpendicular to v in $\mathbb{R}^2$; *i.e.* if $v = (v_1, v_2)$ then $v^\perp = (-v_2, v_1)$ which implies that v is rotated in positive orientation by $\frac{\pi}{2}$. In what follows $\mathrm{Img}(f)$ denotes the image of f.

gradients of f and g have identical sign for all points in $M(z)$, then g differs from f on M:

$$\exists s \in \{-1, 1\} : \operatorname{sgn}\left(\nabla g(m) \cdot \nabla f(m)^{\perp}\right) \equiv s \forall m \in M(z) \Longrightarrow g(v) \neq g(z) \forall v \in M(v) \setminus \{z\}.$$

Proof. Let $z = (z_1, z_2), v = (v_1, v_2) \in M(z)$. Without loss of generality let $z_1 < v_1$. If $z_1 = v_1$ then by Rolle's theorem an optimum would exist on $\{(z_1, y|y \in N\}, N = [z_2, v_2] \cup [v_2, z_2]$ in contradiction to the strict monotonicity[10]. By the theorem of implicit functions a function $j_0(t; z) : [0, 1] \rightarrow \{(y_1, y_2)|y_1 \geq z_1\}$ exists such, that $f(j_0(t)) = f(z) \forall t \in [0, 1]$. Note, that for a given neighborhood of z, j_0 is unique. Now, choose $j_{i+1}(t; m_{i+1})$ with $m_{i+1} = \{(\sup\{m|(m, w) \in \operatorname{Img}(j_i)\}, w)\,|w \in N\} \cap M(z)$, *i.e.* we choose $m_{i+1} = (m_{(i+1)1}, m_{(i+1)2})$ as the rightmost point in $\operatorname{Img}(j_i)$. It follows that $m_{S1} := \sup\{m_{i1}\} = v_1$. Otherwise we can apply the implicit function theorem for $x = m_{S1}$ and find a function $j^{(0)}(t; (m_{S1}, m_{S2}))$ extending the sequence of fuctions j_i, a contradiction, with m_{S2} chosen appropriately. In conclusion we can construct a function $j(t) : [0, 1] \rightarrow \mathbb{R}^2$ such, that $j(0) = z, j(1) = v, j(t) \in M(z)$. Now, choose $t_0 \in [0, 1]$ arbitrarily. By the definition of the gradient $j(t_0 + \epsilon), \epsilon > 0$ can be approximated by $\nabla f|_{j(t_0)}^{\perp}$ with an error in L_2-norm that is $o(\| \epsilon \nabla f|_{j(t_0)}^{\perp} \|^{-1})$. With remark 6.3 g is now strictly monotonic in a neighborhood of $j(t_0)$ and positive ϵ. Depending on s, g is increasing or decreasing. Since by assumption s is constant for all $t \in [0, 1]$. $g \circ j$ is a strictly monotonic function and $g(j(0)) \neq g(j(1))$.

□

■ **Lemma 6.5.** *The likelihood is strictly monotonically increasing in parameter $p^\star$ for fixed β and strictly monotonically increasing for parameter β for fixed $p^\star$ for the outcome of only affected offspring. For only unaffected offspring the negative likelihood has the same properties.*

[10] Alternatively, because of the monotonicity all gradients have positive entries. This implies that each curve in the level set never parallels the axes.

Proof. We denote the likelihood for the events of unaffected offspring with $l_0(\beta, \eta)$ and with $l_k(\beta, p^\star)$ for affected offspring. Let us consider unobserved parents and $k \geq 2$ offspring. For our special assumption, that all family members carry the same diplotypes at L on which we condition for likelihood component L_2, the number of affected offspring is the only observable unit, if we assume parents to be missing. WLOG we can consider this marginal likelihood as identifiability of this likelihood implies identifiability of the full likelihood (argument given in the foot note of page 51):

$$l_k(p^\star, \beta) = \sum_{i=0}^{4} \left(\left(\sum_{j=0}^{2} e_{ij} f(j; \beta) \right)^k \binom{4}{i} p^{\star i} (1 - p^*)^{4-i} \right),$$

with constants e_{ij}, $0 \leq e_{ij} \leq 1 \forall i, j$, $\sum_j e_{ij} = 1$ and f the penetrance function. i enumerates the number of predisposing alleles and the binomial coefficient reflects the distribution of these alleles across the slot of grand-paternal chromosomes. For a given parental configuration the coefficients $c_i := \left(\sum_{j=0}^{2} e_{ij} f(j; \beta)\right)$ is the probability of offspring i being affected which is a mixture of the probabilities for the genotypes $\{0, 1, 2\}$. It remains to be shown that the c_i are ordered in size. First, we have $c_0 = f(0, \beta) < f(2, \beta) = c_4$ by assumption.

Recall, that i counts the number of predisposing alleles in both parents. This implies, *e.g.* that for $i = 0$ only genotype 0 can occur in offspring, *i.e.* $e_{00} = 1$. Table 6.2 enumerates the probabilities which can easily be inferred from independent Mendelian segregation. For example, e_{20} is derived as follows: in two out of 6 cases the two predisposing alleles are present in a single parent; in the other cases the probability is $\frac{1}{2}$ that no allele is transmitted to the offspring. From this table it is clear that $c_0 \leq c_1 \leq c_2 \leq c_3 \leq c_4$ by taking into account the assumptions on the penetrance function.

For all unaffected offspring, obviously, c_i is replaced by $\left(\sum_{j=0}^{2} e_{ij}(1 - f(j; \beta))\right)$ and therefore, by lemma 6.2 and the previous elaboration, l_0 is strictly monotonically decreasing both in $p^\star$ and β and therefore $-l_0$ is strictly increasing in both parameters.

□

Table 6.2: Enumeration of the factors e_{ij} in lemma 6.5, see text. i is varied across columns, j across rows.

	0	1	2	3	4
0	1	$\frac{1}{2}$	$\frac{1}{6}$	0	0
1	0	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{2}$	0
2	0	0	$\frac{1}{6}$	$\frac{1}{2}$	1

With lemma 6.5 it only remains to check the condition on the scalar products of gradients of the likelihood for observations of affected/unaffected offspring to show identifiability with lemma 6.4. As it turns out this condition is hard to check symbolically. Instead, we pursue a numeric approach by inspecting several informative plots to check assumptions for lemma 6.4. In conclusion, we have the following corollary.

■ **Corollary 6.6.** *Given that the scalar product of the gradients of likelihood l_0 and l_k are negative for all parameter values, i.e.*

$$\nabla - l_0(p^\star, \beta)^\perp \cdot \nabla l_k(p^\star, \beta) > 0 \, \forall(\beta, p^*),$$

the likelihood for random sampling and unobserved parents is identifiable.

□

This corollary is the application of lemma 6.4 to the likelihoods l_0 and l_k.

We now restate the likelihoods for only affected (l_k) and only unaffected offspring (l_0) and a general penetrance function. When all offspring are affected, we get for l_k:

$$l_k(p^\star, \beta) = \sum_{i=0}^{4} \left(\left(\sum_{j=0}^{2} e_{ij} f(j; \beta) \right)^k \binom{4}{i} p^{\star i} (1 - p^*)^{4-i} \right).$$

For k unaffected offspring we have:

$$\begin{aligned}
l_0(p^*,\beta) &= \sum_{i=0}^{4}\left(\left(\sum_{j=0}^{2} e_{ij}(1-f(j;\beta))\right)^k \binom{4}{i} p^{*i}(1-p^*)^{4-i}\right)\\
&= \sum_{i=0}^{4}\left(\left(1-\sum_{j=0}^{2} e_{ij}f(j;\beta)\right)^k \binom{4}{i} p^{*i}(1-p^*)^{4-i}\right)\\
&= \sum_{i=0}^{4}\left(\left(\sum_{n=0}^{k}\binom{k}{n}(-1)^n\left(\sum_{j=0}^{2} e_{ij}f(j;\beta)\right)^{k-n}\right)\binom{4}{i} p^{*i}(1-p^*)^{4-i}\right)\\
&= 1-\sum_{i=0}^{4}\left(\left(\sum_{n=1}^{k-1}\binom{k}{n}(-1)^n\left(\sum_{j=0}^{2} e_{ij}f(j;\beta)\right)^{k-n}\right)\binom{4}{i} p^{*i}(1-p^*)^{4-i}\right)+l_k(\beta,p^*).
\end{aligned}$$

Note, that for symbolic treatment we can subtract l_2 from l_0, preserving monotonicity characteristics and reducing the polynomial complexity of l_0 by one. If, for example, we consider level sets of l_2, joint level sets of l_2 and l_0 are preserved for $l_0 + l_2$.

No symbolic solution can be found for $\nabla - l_0(p^\star,\beta)^\perp \cdot \nabla l_k(p^\star,\beta) > 0$ as is apparent from the above formulas. However, numeric evaluation of this expression is straightforward. The criterion involves only likelihoods for two outcomes which makes it much easier to evaluate and interpret as compared to, say, numerically evaluating the empirical Fisher information on a close grid of parameter values for representative data sets.

In summary, we have proven that a criterion on the gradients of certain likelihood functions is equivalent to the identifiability of the full likelihood. A symbolic treatment of this criterion is not possible for the logistic penetrance function that is used in this thesis such that we resort to a numeric evaluation of the criterion given in corollary (6.6) in the following section.

6.3. **Checking identifiability conditions for a concrete example.** In this example we consider $k = 2$ offspring. The liklihood contributions for 2 affected and 2 unaffected

Figure 6.1: Contour plot of product of gradients of likelihood for two offspring for only affecteds and no affecteds. β is plotted on the x-axis and p^* on the y-axis.

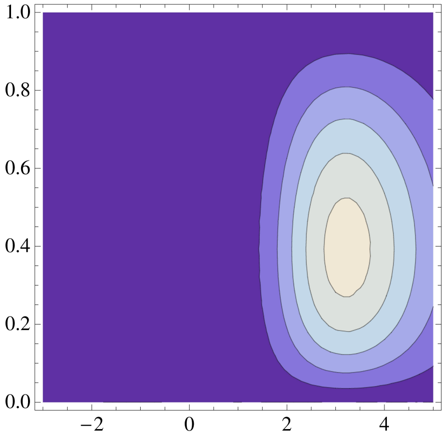


Figure 6.2: 3D plot of product of gradients of likelihood for two offspring for only affecteds and no affecteds. β ranges from -3 to 5 (x axis, lower left) and p^* ranges from 0 to 1 (y axis, right).

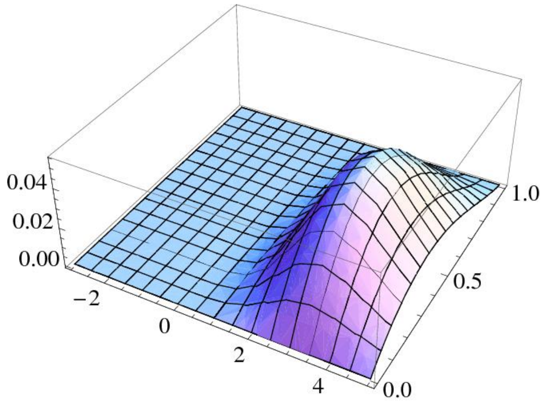
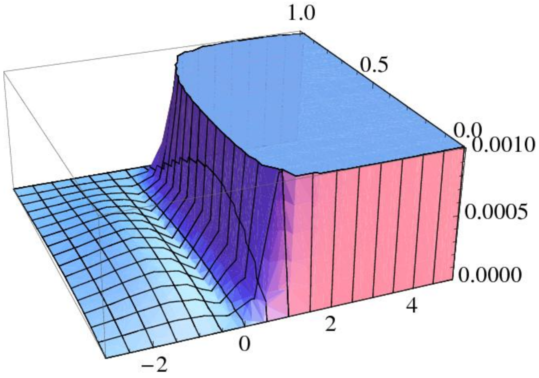


Figure 6.3: 3D plot of product of gradients of likelihood for two offspring for only affecteds and no affecteds. β ranges from -3 to 5 (x axis, bottom) and p^* ranges from 0 to 1 (y axis, top right). z axis is cut off at 0.001.



offspring are then respectively given by:

$$
\begin{aligned}
l_2(p^\star,\beta) &= \sum_{i=0}^{4}\left(\left(\sum_{j=0}^{2} e_{ij} f(j;\beta)\right)^2 \binom{4}{i} p^{\star i}(1-p^*)^{4-i}\right)\\
l_0(p^*,\beta) &= 1-2\sum_{i=0}^{4}\left(\left(\sum_{j=0}^{2} e_{ij} f(j;\beta)\right)\binom{4}{i}\eta^i(1-\eta)^{4-i}\right)+l_2(\eta,\beta).
\end{aligned}
$$

Figures 6.1-6.3 show plots of the term $\mathbb{L} := \nabla l_2(\beta, p^\star)\nabla(-l_0(\beta, p^\star))^\perp$ for two offspring. μ is set to -2, the mode of inheritance is dominant and $p^\star$ is plotted in the range $(0, 1)$ and β ranges over $(-3, 5)$. The 3D plot in figure 6.2 and 6.3 differ in the range of $z = \mathbb{L}(\beta, p^*)$ and the viewing point. Figure 6.1 shows a contour plot of $\mathbb{L}(\beta, p^\star)$ in $p^\star$ and β. Dark shades represent small values and light shades larger values. From this plot is appears that $\mathbb{L}$ has a single maximum at $(p^\star, \beta) \approx (0.4, 3.5)$. Plotting $\mathbb{L}^2$ can be used to check for negative regions of $\mathbb{L}$. However, the plot of $\mathbb{L}^2$ looks very similar to figure 6.1 (figure not shown). The second graphical check is performed in figure 6.2 that represents a three-dimensional plot of $\mathbb{L}$. Here, the non-negativity and the single maximum is corroborated. To further explore the behaviour of $\mathbb{L}$, figure 6.3 again shows a three-dimensional representation, this time plotting $\min\{0.001, \mathbb{L}\}$ to allow for an inspection of the flat region for $\beta < 0$. It is visible that $\mathbb{L} = 0$ for $\beta = 0$. For $\beta < 0$ $\mathbb{L}$ is positive again, albeit the function is smaller by several orders of magnitude as compared to the region $\beta > 0$.

6.3.1. *The case $\beta = 0$:* For $\beta = 0$, the likelihood becomes independent of locus $L^\star$, since there is no effect of any genetic locus. In this case the parameter p^* is undefined and the likelihood depends only on η. For random sampling the likelihood terms for affected and unaffected indivduals therefore become collinear and thereby $\nabla l_2(\beta, p^\star)\nabla(-l_0(\beta, p^\star))^\perp = 0$. This case is excluded by assumption (A5).

6.3.2. *The case $\beta \neq 0$:* To assure global identifiability, beta has to be chosen either $\beta > 0$ or $\beta < 0$, since only then the assumptions of lemma 6.4 are fulfilled. This however reflects the fact that the joint distribution of haplotypes is based on an arbitrary assignment of alleles at $L^\star$. For example, if we consider one observed and one unobserved locus, p^* corresponding to some ξ_d can be considered the conditional allele frequency of allele 1 at loci $L^\star$ given that allele 1 at L_1 was observed. By a reassignment of allele naming, however, p^* might likewise parametrize the dependency between allele 1 at L_1 and allele 2 at $L^\star$. The model does not make this distinction explicit. Implicitly, if β is estimated to be positive, ξ_d measures dependency with a positively associated allele at $L^\star$ and the other allele is therefore negatively associated. Thus, without loss of generality, β can be restricted to be positive on account of the possibility to rename alleles.

By use of these numeric techniques a particular likelihood can be checked for identifiability.

6.4. **Consistency of MLEs.**

■ **Lemma 6.7.** *The MLE $\hat{\theta}$ of likelihood $L(\theta; Y, G)$ is consistent for true parameter θ_0:*

$$\hat{\theta} := \arg\max_{\theta} L(\theta; G) \xrightarrow{P} \theta_0$$

Proof. To show consistency of the MLE, a sufficient criterion is identifiability of the likelihood and existence of the Fisher-Information matrix ([56], ch. 7), which asymptotically converges to the inverse of the covariance matrix of the parameter estimates. It therefore remains to show existence of the Fisher-Information.

To this end, we switch to parametrization 1 and form (5.8) of the likelihood. To derive derivatives with respect to η^* with β fixed, we write the likelihood as follows:

$$L(\eta^*) \quad = \sum_{\mathbf{h_1^{*1}}|\mathbf{h_1^{*2}}=\mathbf{g_1}} c_{h^*} \prod_{j=1}^{2} P(H^*_{1j} = h^*_{1j}), \tag{6.4}$$

where penetrance terms $P(\mathbf{Y}|\mathbf{G}^* = \mathbf{g}^*)$ and Mendelian constants are collapsed into c_{h^*}. As any factor $P(H^*_{1j} = h^*_{1j})$ is a polynomial of degree 2 in $\eta^*_1, ..., \eta^*_{2^K}$ (section 2.2.3), we can rewrite $L(\eta^*)$ as follows:

$$L(\eta^*) \quad = \quad \sum_{|\mathbf{j}|=4} \tilde{c}_{\mathbf{j}} \prod_{i=1}^{2^K} \eta_i^{*j_i},$$

where $\mathbf{j} = (j_1, ..., j_{2^K}), j_i > 0$ and $|\mathbf{j}| = \sum j_i$. Therefore, the score is given by:

$$\frac{\partial}{\partial \eta^*_k} \log(L(\eta^*)) \quad = \quad \frac{\sum_{|\mathbf{j}|=3} \hat{c}_{\mathbf{j}} \prod_{i=1}^{2^K} \eta_i^{*j_k}}{\sum_{|\mathbf{j}|=4} \tilde{c}_{\mathbf{j}} \prod_{i=1}^{2^K} \eta_i^{*j_i}},$$

where $\hat{c}_{\mathbf{j}}$ absorbs exponents resulting from taking derivatives. For the second derivative we have:

$$\frac{\partial^2}{\partial\eta_k^*\partial\eta_l^*}\log(L(\eta^*)) \quad = \quad \frac{\sum_{|\mathbf{j}|=7}\bar{c}_{\mathbf{j}}\prod_{i=1}^{2K}\eta_i^{*j_k}}{(\sum_{|\mathbf{j}|=4}\tilde{c}_{\mathbf{j}}\prod_{i=1}^{2K}\eta_i^{*j_i})^2},$$

where factors are again changes to $\bar{c}_{\mathbf{j}}$. Taking the derivative with respect to β and treating terms η^* constant we write the likelihood as:

$$L(\beta) \quad = \quad \sum_{\mathbf{x}^*} d_i \prod_{j=1}^{k} P(Y = 1|X^* = x_j^*),$$

where the $\mathbf{x}^* = (x_1^*, x_k^*)$ iterates all score combinations as corresponding to genotypes at the disease locus. The score function is given by:

$$\frac{\partial}{\partial\beta}\log L(\beta) \quad = \quad \frac{\sum_{\mathbf{x}^*} d_i \frac{\partial}{\partial\beta}\prod_{j=1}^{k} P(Y = 1|X^* = x_j^*)}{\sum_{\mathbf{x}^*} d_i \prod_{j=1}^{k} P(Y = 1|X^* = x_j^*)}$$

From this, it follows that $\frac{\partial^2}{\partial\beta^2}\log L(\beta)$ is a rational function in $\frac{\partial^2}{\partial\beta^2}P(Y = 1|X^* = x^*)$, $\frac{\partial}{\partial\beta}P(Y = 1|X^* = x^*)$ and $P(Y = 1|X^* = x^*)$ for every possible score x^*. We therefore need the penetrance function to be differentiable twice. For example, the logistic penetrance function has derivative:

$$\begin{aligned}
& \frac{\partial}{\partial\beta}\frac{\exp(\mu + \beta x^*)}{1 + \exp(\mu + \beta x^*)} \\
= \quad & \frac{\exp(\mu + \beta x^*)x^*(1 + \exp(\mu + \beta x^*)) - \exp(\mu + \beta x^*)^2 x^*}{(1 + \exp(\mu + \beta x^*))^2} \\
= \quad & \frac{\exp(\mu + \beta x^*)x^*}{(1 + \exp(\mu + \beta x^*))^2},
\end{aligned}$$

and second derivative

$$\begin{aligned} & \frac{\partial^2}{\partial \beta^2} \frac{\exp(\mu + \beta x^*)}{1 + \exp(\mu + \beta x^*)} \\ = \quad & \frac{\partial}{\partial \beta} \frac{\exp(\mu + \beta x^*) x^*}{(1 + \exp(\mu + \beta x^*))^2} \\ = \quad & \frac{\exp(\mu + \beta x^*) \left\{x^*(1 + \exp(\mu + \beta x^*))\right\}^2}{\left\{1 + \exp(\mu + \beta x^*)\right\}^4}. \end{aligned}$$

The existence of mixed derivatives $\frac{\partial^2}{\partial \eta_i^* \partial \beta}$ follows from that fact, that the $\bar{c}_{\mathbf{j}}$ are polynomials in $P(Y = 1 | X^* = x^*)$. As second derivatives are finite for all permissible parameters and the distribution is discrete with finite support, expectations are finite as well and the Fisher information exists.

□

6.5. **Identifiability of the random effects models.** I do not show identifiability for the model with $\sigma_a \neq 0$, but give arguments making it plausible. Additional constraints on samples become apparent.

In the foregoing proof identifiability of parameters (β, p^*) was shown conditional on a fixed diplotype. For a single family, this implies that a consistent estimate $(\hat{\beta}, \hat{p}^*_F)$ can be obtained for the asymptotic sample $N = 1, n_1 \to \infty$, *i.e* we sample a single family for which offspring size tends to ∞. Here, $\hat{p}^*_F$ would be the frequency of the predisposing allele A in the parents and would not tend to $\hat{p}^*$, the population frequency of allele A in general. However, β would be estimated consistently even in the single family. As a result the variability in estimates $\hat{\beta}$ across families as $N \to \infty$ would asymptotically reflect the influence of the random effect that would therefore be identifiable.

Adding parameter σ_a has implications for family structure requirements to achieve identifiability. We noted before, that at least two offspring are needed to identify $p^\star$ and β, if no parents are observed. In this instance the estimation of $p^\star$ and β relied solely on the phenotype distribution of offspring, conditional on a fixed diplotype in the family. However, the bivariate Bernoulli distribution of binary phenotypes of two offspring is fully characterized by two parameters as offspring are exchangeable and marginal probabilities are therefore identical. This implies that a maximum of two parameters can be identified with two offspring and therefore at least three offspring are needed to identify the full set of parameters (β, p^*, σ_a).

7. Bayesian approach

There are three major motivations for employing a Bayesian framework for this model: (1) multiple latent entities (*e.g.* haplotypes involving the disease locus) can be naturally modeled in the Bayesian framework; (2) the number of parameters increases exponentially with the number of observed loci which makes optimization challenging; (3) as we expect applications of these methods in fine mapping steps, *a-priori* knowledge, that a given region is associated with a disease outcome can be incorporated.

Bayesian inference is based on the posterior distribution of parameters $\theta = (\eta^\star, \beta)$, parametrization 1, section 5.2.1, given the observed data. The posterior distribution is given for this likelihood as follows:

$$P(\theta|Y,G) = \frac{P(Y,G|\theta)P(\theta)}{\int_\theta P(Y,G|\theta)P(\theta)} \propto P(Y,G|\theta)P(\theta) \tag{7.1}$$

7.1. Data augmentation.

$P(Y,G|\theta) = P(Y|G,\theta)P(G|\theta)$ can be computed by marginalizing over the disease locus. We denote the diplotypes at observed and disease locus with $H^\star$ and assume that $h^{\star 1}|h^{\star 2} = g$ is the set of diplotypes compatible with observed genotypes in the whole sample. We therefore get:

$$\begin{aligned}
P(Y,G=g|\theta)P(\theta) &= \sum_{h^{\star 1}|h^{\star 2}=g} P(Y,G=g,H^\star = h^\star|\theta)P(\theta) \\
&= \sum_{h^{\star 1}|h^{\star 2}=g} P(Y|G,H^\star = h^\star,\theta)P(G|H^\star = h^\star,\theta)P(H^\star = h^\star|\theta)P(\theta) \\
&= \sum_{h^{\star 1}|h^{\star 2}=g} P(Y|H^\star = h^\star,\theta)P(H^\star = h^\star|\theta)P(\theta) \\
&= \left\{ \sum_{h^{\star 1}|h^{\star 2}=g} P(Y|G^\star = g^\star,\theta)P(H^\star = h^\star|\theta) \right\} P(\theta). \qquad (7.2)
\end{aligned}$$

If $H^\star$ were observed, $P(Y,G|\theta)$ would be straightforward to compute. The idea of data augmentation is to impute realizations $H^\star$ from the conditional predictive distribution

$P(H^\star|\theta, G, Y)$ and then use the full data $(Y, G, H^\star)$ to straightforwardly compute likelihood 7.2. The posterior distribution of θ is approximated by averaging the imputed posterior probabilities for m realizations of $H^{\star(i)}$ (7.1) [98, 61] as:

$$P(\theta|Y,G) = \frac{1}{m}\left(\sum_{i=1}^{m} P(\theta|Y, H^{\star(i)})\right). \tag{7.3}$$

The data augmentation algorithm is iterative and involves drawing θ from the approximated posterior and generating new $H^\star$ values for a given realization of θ in each step (iterate between computing (7.2) and (7.3)). The corresponding densities are given in section 7.5.

7.2. **Collapsing.** Data augmentation requires to alternatingly draw parameters and missing data. This procedure can be optimized by establishing a predictive distribution $P(H^{\star(i+1)}|H^{\star(i)}, Y)$, defined by integrating over parameters for the missing data [61]:

$$P(\beta, Y, G, H^\star) = \int_{\eta^*} P(\beta, \eta^*, Y, G, H^\star) d\eta^*.$$

Collapsed samplers have previously been proposed to solve the phase ambiguity for reconstructing haplotypes from observed genotypes [73, 95, 94]. In these instances, haplotypes are updated for each individual in the sample by a Gibbs sampler: $P(H^{\star(i+1,j)}|H^{\star(i,-j),Y}), j = 1, ..., n$, *i.e.* new haplotypes are predicted for each person based on predicted haplotypes for the remainder of the sample. Approaches differ by the choice of prior distributions for haplotype frequencies. For example, Niu et al. 2002 [73] assume a Dirichlet prior and derive the following Gibbs sampler:

$$P(h_i^\star|H_{(-i)}^\star, G) = P((h_{i1}^\star, h_{i2}^\star)|H_{(-i)}^\star, G) \propto (n_1 + \gamma_1)(n_2 + \gamma_2).$$

Here, γ_i are the parameters of the Dirichlet distribution and n_1 and n_2 denote haplotype counts for $h^\star_{i1}$ and $h^\star_{i2}$ in the sample $H^\star_{(-i)}$, respectively. If the n_is are large compared to the γs, haplotypes are drawn roughly according to their sample frequency. Stephens et. al [95] do not specify an *a-priori* distribution but rather directly give a collapsed sampler $P(H^{\star(i+1,j)}|H^{\star(i,-j)})$ (section 3.4.3). The *a-priori* distribution is therefore implicit and the justification of the sampler comes from the practical perspective that it outperforms other algorithms in real data sets.

To apply a collapsed sampler for haplotype frequencies in our problem, two modifications are necessary: (1) the predictive distribution has to include the phenotype, *i.e.* $H^{\star(i+1,j)}$ depends on Y_j: $P(H^{\star(i+1,j)}|H^{\star(i,-j)}, Y_j)$ and (2) the sampling unit is the nuclear family. Dependencies among family members have to be accounted for in independent draws from the predictive distribution. It should be noted, that sampling of $H^\star$ is conditional on G, which implies that allele frequencies are fixed to sample frequencies during the updating. This is further discussed in section 7.6.

7.3. **Sampling strategy.** Estimation of the posterior distribution is performed by defining a Monte Carlo Markov Chain (MCMC) [23]. Haplotype frequencies are collapsed and haplotypes are updated from a predictive distribution. The penetrance parameter β is updated by assuming a normal prior distribution. Haplotype probabilities are updated by their sample frequencies. I will now define the model in detail and describe the updating for a collapsed sampler for the likelihood of this thesis.

7.4. **Prior distributions and deterministic relationships.** Recall that parametrization 1 (section 5.2.1) is used for likelihood (5.4) (page 40). This implies that haplotypes are assumed to be distributed as a multinomial. The distributions in the model are for a family with k members:

$$\begin{aligned} H_{ij}^{\star k} &\sim \text{Mult}(1, \eta_1^*, ..., \eta_{2^{K+1}}^*) \\ Y_i &\sim B(1, p_i), i \in \{1, ..., k\} \\ \text{logit}(p_i) &= \mu + \beta X_i^\star. \end{aligned}$$

Note that $P(Y_i|H_i^\star) = P(Y_i|G_i^\star)$. $X_i^\star$ is the score associated with $(H_{i1}^\star, H_{i2}^\star)$ (*c.f.* section 2.3, page 15). The following distributional assumptions model the prior probabilities:

$$\begin{aligned} \beta &\sim N(\beta_0, \sigma_\beta^2) \\ \eta^\star &\sim \text{Dirichlet}(\gamma_1, ..., \gamma_m) \end{aligned}$$

$\beta_0, \sigma_\beta^2, \gamma_1, ..., \gamma_m$ are hyperparameters of the model. We follow an empirical Bayes approach by choosing parameters leading to mildly informative prior distributions. The specific choice is given in the simulation section.

7.5. **Densities for sampling distributions.** All sampling densities can be defined through the joint augmented likelihood:

$$P(\theta, H^\star, G, Y) = P(Y|H^\star, \beta)P(G|H^\star, \beta, \eta^*)P(H^\star|\beta, \eta^*)P(\theta)$$

$$\begin{aligned} &= P(Y|H^\star, \beta)P(G|H^\star)P(H^\star|\eta)P(\theta) \\ &= P(Y|G^\star, \beta)P(H^\star|\eta^*)P(\theta)I\{h_1^\star|h_2^\star = g\}, \end{aligned}$$

with $I\{h_1^\star|h_2^\star = g\}$ being the indicator whether $H^\star$ is compatible with G. It is assumed that $P(H^\star|\beta, \eta^*)$ is independent of β. This is an additional assumption to those mentioned in section 4.2. This assumption means that the dependency of H and $G^\star$ is independent of β. In practical terms this implies that the genomic regions investigated are chosen independently of β, which is a very plausible assumption. Recall, that the sampling unit is a nuclear family, and therefore haplotypes have to be compatible with all observed genotypes in a family. The algorithmic approach is described below (section 7.6). I make the notational assumption that $H^\star$ is always restricted to be compatible with genotypes in the following.

The Markov Chain proceeds by iteratively drawing from the following distributions:

$$\begin{aligned} P(H^\star|\theta, G, Y) &\propto P(H^\star|H^{\star(-1)})P(Y|H^\star, \beta) \\ P(\beta|H^\star, G, Y) &\propto P(\beta)P(Y|H^\star, \beta)P(H^\star). \end{aligned}$$

The posterior distribution $P(H^\star|G, Y)$ is obtained as the collapsed $\int P(H^\star|\theta, G)P(\theta)d\theta$ followed by a Metropolis-Hastings step to account for the dependency on Y. β has to be updated by means of a Metropolis step, since the full conditional distribution for β does not have a closed form.

7.6. **Updating** $H^\star$**.** For each family diplotypes are updated jointly for all members as follows:

- Draw a new family from the proposal distribution $P(H^\star_i|H^\star_{-i})$,
- Use a Metropolis-Hastings step to accept or reject these diplotypes.

Joint updating of haplotypes of members of a family is done by iterating family members, each time conditioning on already updated members. First, a single parent is drawn according to the predictive distribution $P(h^\star_i|H^\star_{(-i)}, G) = P((h^\star_{i1}, h^\star_{i2})|H^\star_{(-i)}, G) \propto (n_1 + \gamma_1)(n_2 + \gamma_2)$. This is the same distribution as the one used for the collapsed sampler to solve phase ambiguity (section 3.4.4). Here, it serves as the proposal distribution in the Metropolis-Hastings step. n_1 and n_2 are haplotype counts in all parents. Next, diplotypes of the second parent is drawn conditional on the diplotype of the first parent and offspring genotypes. Finally, offspring diplotypes are drawn according to Mendelian inheritance in accordance with genotypes. This Gibbs sampler is the proposal distribution of the Metropolis-Hastings step. The family is accepted with probability α (denoting the proposed diplotypes with $H^\star_N$ and the old diplotypes with $H^\star_O$):

$$\alpha = \min\left(1, \frac{P(Y, H_N^\star; \beta)P(H_O^\star)}{P(Y, H_O^\star; \beta)P(H_N^\star)}\right),$$

which is evaluated for the single family. One advantage of the collapsed sampling approach is that the likelihood is evaluated for complete data. The likelihood for a single family has the following form:

$$\begin{aligned} P(Y_i, H_i^\star) &= \prod_{j=1}^{n_i} P(Y_{ij}|H_{ij}^\star = h_{ij}^\star)P(H_{ij}^\star = h_{ij}^\star|h_{iM}^\star, h_{iP}^\star) \\ &\times \prod_{j\in\{M,P\}}^{n_i} P(Y_{ij}|H_{ij}^\star = h_{ij}^\star)P(H_{ij}^\star = h_{ij}^\star) \\ &= \underbrace{\prod_{j\in F}^{n_i} P(Y_{ij}|H_{ij}^\star = h_{ij}^\star)}_{A} \underbrace{\prod_{j\in\{M,P\}}^{n_i} P(H_{ij}^\star = h_{ij}^\star)\prod_{j=1}^{n_i} P(H_{ij}^\star = h_{ij}^\star|h_{iM}^\star, h_{iP}^\star)}_{B}, \end{aligned} \tag{7.4}$$

with $F = \{M, P, 1, ..., n_i\}$. This representation is considerably simpler than the likelihood given in section 5.2, since it factorizes into phenotype part A and the diplotype part B. This can be used to make the updating process computationally efficient, since computing B is equivalent to computing the likelihood of the proposal distribution. Computation of the joint distribution only involves computing B and adding the terms on the logarithmic scale.

7.7. **Updating of β.** β can be updated by a Metropolis-step, since the proposal distribution $N(\beta_O, \sigma_\beta^2)$ is symmetric, where β_O is the value of β in the previous iteration and σ_β^2 its variance. I accept a new update with probability α:

$$\alpha = \min\left(1, \frac{P(Y, H^\star; \beta_N)}{P(Y, H^\star; \beta_O)}\right),$$

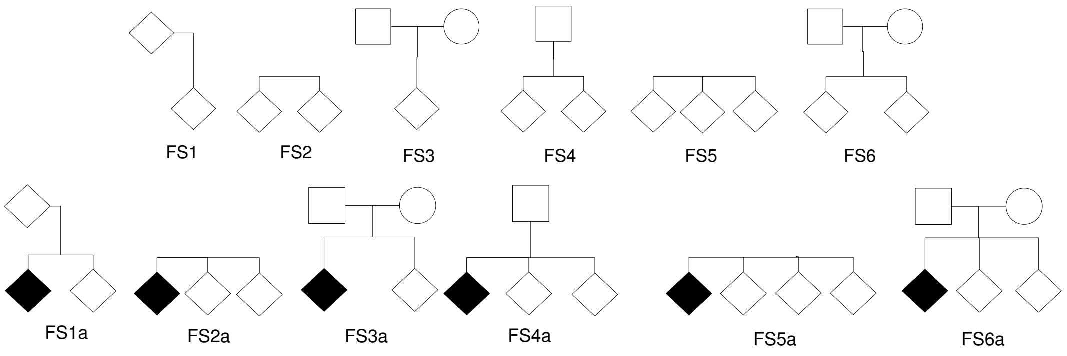
An example of a single run of the MCMC sampler is shown in figure 8.10 on page 86.

8. Simulation study

In this chapter the performance of the model is assessed in terms of estimation and testing by means of simulations.

8.1. **Parameter estimation and comparison of family structures.** Parameters were estimated from simulated data in order to evaluate the impact of ascertainment, sample size, family structure and model misspecification with respect to mode of inheritance. In all simulations, a parametric model is used to generate genotypes and phenotypes according to true parameter values. Mean squared errors and standard deviations of parameter estimates are computed. Also the impact of distributing a fixed number of individuals among different family structures is considered, which is an important aspect for study designs. The family structures considered are displayed in figure 8.1. In ascertained samples, ascertainment is always on at least one affected offspring.

Figure 8.1: Abbreviations for pedigrees of family structures considered in the simulations. The first row shows randomly sampled families. Pedigrees are organized by a single row per generation, connecting spouses by straight lines, which connect to offspring by a vertical line. Circles are females, squares are males and diamonds designate unknown sex. Families ascertained on at least one affected offspring are shown in the second row.



A complete list of simulations is given in appendix C. Table C.1 (page 122) lists all parameter combinations considered in the single locus case together with table and page numbers. Table C.29 (page 168) gives the same information for the two locus case.

8.2. **Parameter estimation under the null hypothesis.** Parameter estimates under the null hypothesis are summarized in table 8.1. For complete iterations of scenarios refer to appendix C (table C.2, page 123 through table C.28, page 167). Parameter estimates are accurate for random sampling (table 8.1). Under misspecification of the mode of inheritance the additive model provides robust estimations under many circumstances. Recall, that the modes of inheritance differ in how a score is assigned to the causal locus (section 2.3.1) which then linearly changes disease probably from a baseline on a *logit* scale. The additive model uses allele count of the predisposing allele as score, *i.e.* being homozygous is more relevant than being heterozygous. The dominant model assigns the same score to any genotype that contains the predisposing allele, whereas the recessive model only scores homozygous genotypes for the predisposing allele. Returning to robustness, the estimates for a data set simulated under the dominant model are roughly as accurate as estimates under the correct model, as measured by mean squared error (MSE; table 8.1). With random sampling, the recessive model can be problematic. In this case too few affected individuals are present in the samples (sample size $\sim$ 100 families). Interestingly, fitting an additive or dominant model can give more robust estimates than fitting the recessive model. This pattern also occurs in ascertained samples (table 8.2). Under misspecification the correlation R_1 is estimated accurately, however, a bias is introduced in parameters $p^\star$ and β. For example, for recessive data, the dominant and additive models overestimate $p^\star$ and β, although family structures FS6/FS6a seem to result in fairly robust estimates even under misspecification.

8.3. **Ascertainment.** For small allele frequencies (0.1) and moderate effect sizes ($\beta = 2$), the number of affecteds can be low in random samples. Table 8.3 shows the effect of ascertainment on family structures FS6/FS6a. MSEs are consistently lower for ascertained samples. Like under the null hypothesis it seems to be better to fit recessive data with either the dominant or the recessive model.

8.4. **Family structure.** Family structure is an important issue in study design. For example, it can be advantageous to sample sibs only, since they have a more homogeneous

Table 8.1: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.0$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	150	0.20 (0.02; 0.00)	-0.00 (0.34; 0.12)	0.28 (0.23; 0.06)	1.52 (1.12; 1.49)
FS2	150	0.20 (0.02; 0.00)	0.01 (0.37; 0.14)	0.33 (0.23; 0.07)	1.23 (0.62; 0.98)
FS3	100	0.20 (0.02; 0.00)	-0.00 (0.32; 0.10)	0.30 (0.23; 0.06)	2.10 (1.20; 1.44)
FS4	100	0.20 (0.02; 0.00)	0.01 (0.33; 0.11)	0.28 (0.23; 0.06)	1.62 (0.97; 1.10)
FS5	100	0.20 (0.02; 0.00)	0.01 (0.34; 0.11)	0.32 (0.24; 0.07)	1.49 (0.65; 0.68)
FS6	75	0.20 (0.02; 0.00)	0.00 (0.31; 0.10)	0.31 (0.23; 0.06)	1.93 (0.69; 0.49)
	simulation: dom; estimation: add				
FS1	150	0.20 (0.02; 0.00)	-0.02 (0.28; 0.08)	0.23 (0.22; 0.05)	1.91 (0.95; 0.91)
FS2	150	0.20 (0.02; 0.00)	0.00 (0.34; 0.12)	0.30 (0.23; 0.06)	1.54 (0.66; 0.65)
FS3	100	0.20 (0.02; 0.00)	0.00 (0.25; 0.06)	0.24 (0.19; 0.04)	2.59 (1.26; 1.92)
FS4	100	0.20 (0.02; 0.00)	0.02 (0.25; 0.06)	0.22 (0.19; 0.04)	2.06 (0.84; 0.71)
FS5	100	0.20 (0.02; 0.00)	-0.01 (0.27; 0.08)	0.23 (0.18; 0.03)	1.95 (0.70; 0.49)
FS6	75	0.20 (0.02; 0.00)	-0.01 (0.23; 0.05)	0.22 (0.16; 0.03)	2.49 (0.80; 0.88)

Table 8.2: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.0$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1a	640	0.20 (0.01; 0.00)	-0.00 (0.13; 0.02)	0.15 (0.16; 0.03)	1.74 (0.52; 0.34)
FS3a	480	0.20 (0.01; 0.00)	-0.00 (0.17; 0.03)	0.30 (0.25; 0.07)	2.00 (0.41; 0.17)
FS6a	384	0.20 (0.01; 0.00)	-0.01 (0.17; 0.03)	0.29 (0.23; 0.06)	1.98 (0.37; 0.14)
	simulation: rec; estimation: rec				
FS1a	640	0.20 (0.01; 0.00)	0.10 (0.27; 0.08)	0.51 (0.23; 0.15)	-9.76 ($\geq$ 10; $\geq$ 10)
FS3a	480	0.20 (0.01; 0.00)	-0.00 (0.24; 0.06)	0.37 (0.32; 0.13)	2.87 (3.51; $\geq$ 10)
FS6a	384	0.20 (0.01; 0.00)	-0.00 (0.26; 0.07)	0.34 (0.28; 0.10)	2.59 (3.50; $\geq$ 10)
	simulation: rec; estimation: add				
FS1a	640	0.20 (0.01; 0.00)	0.00 (0.05; 0.00)	0.25 (0.08; 0.01)	4.04 (1.80; 7.39)
FS3a	480	0.20 (0.01; 0.00)	0.00 (0.06; 0.00)	0.59 (0.12; 0.17)	2.59 (0.36; 0.47)
FS6a	384	0.20 (0.01; 0.00)	0.00 (0.06; 0.00)	0.57 (0.09; 0.14)	2.61 (0.28; 0.45)

age structure than nuclear families including parents. Also ethical issues might be relevant, since risk assessment might affect family structures differently. In the simulation study, the overall count of individuals was kept constant and distributed across varying family structures. Examples are shown in tables 8.1 and 8.2. In these cases MSEs for η_1, R_1 and

Table 8.3: Effect of ascertainment on family structures FS6/FS6a. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.5$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS6	300	0.10 (0.01; 0.00)	0.53 (0.16; 0.02)	0.11 (0.05; 0.00)	2.01 (0.46; 0.21)
FS6a	384	0.10 (0.01; 0.00)	0.52 (0.10; 0.01)	0.11 (0.04; 0.00)	2.00 (0.23; 0.05)
	simulation: dom; estimation: rec				
FS6	300	0.10 (0.01; 0.00)	0.45 (0.40; 0.16)	0.12 (0.15; 0.02)	0.46 (2.11; 6.82)
FS6a	384	0.10 (0.01; 0.00)	0.37 (0.30; 0.11)	0.07 (0.10; 0.01)	0.94 (1.87; 4.61)
	simulation: dom; estimation: add				
FS6	300	0.10 (0.01; 0.00)	0.51 (0.12; 0.02)	0.09 (0.04; 0.00)	2.28 (0.43; 0.27)
FS6a	384	0.10 (0.01; 0.00)	0.47 (0.08; 0.01)	0.08 (0.03; 0.00)	2.34 (0.20; 0.16)
	simulation: rec; estimation: dom				
FS6	300	0.10 (0.01; 0.00)	0.39 (0.08; 0.02)	0.33 (0.08; 0.06)	3.85 (3.25; $\geq$ 10)
FS6a	384	0.11 (0.01; 0.00)	0.40 (0.04; 0.01)	0.40 (0.05; 0.10)	2.32 (0.25; 0.17)
	simulation: rec; estimation: rec				
FS6	300	0.10 (0.01; 0.00)	0.61 (0.31; 0.11)	0.17 (0.14; 0.02)	-0.26 ($\geq$ 10; $\geq$ 10)
FS6a	384	0.10 (0.01; 0.00)	0.56 (0.24; 0.06)	0.14 (0.12; 0.02)	2.81 (4.18; $\geq$ 10)
	simulation: rec; estimation: add				
FS6	300	0.10 (0.01; 0.00)	0.42 (0.07; 0.01)	0.31 (0.07; 0.05)	4.67 (3.72; $\geq$ 10)
FS6a	384	0.11 (0.01; 0.00)	0.44 (0.04; 0.00)	0.36 (0.04; 0.07)	2.83 (0.37; 0.83)
	simulation: add; estimation: dom				
FS6	300	0.10 (0.01; 0.00)	0.50 (0.14; 0.02)	0.13 (0.07; 0.01)	1.80 (0.56; 0.35)
FS6a	384	0.10 (0.01; 0.00)	0.43 (0.08; 0.01)	0.09 (0.05; 0.00)	1.93 (0.32; 0.11)
	simulation: add; estimation: rec				
FS6	300	0.10 (0.01; 0.00)	0.51 (0.42; 0.18)	0.14 (0.16; 0.03)	0.34 (2.01; 6.79)
FS6a	384	0.10 (0.01; 0.00)	0.57 (0.34; 0.12)	0.12 (0.12; 0.01)	0.45 (1.43; 4.43)
	simulation: add; estimation: add				
FS6	300	0.10 (0.01; 0.00)	0.52 (0.13; 0.02)	0.11 (0.05; 0.00)	1.99 (0.50; 0.25)
FS6a	384	0.10 (0.01; 0.00)	0.51 (0.08; 0.01)	0.11 (0.04; 0.00)	1.99 (0.26; 0.07)

$p^\star$ do not vary substantially between family structures. For β, FS6 seems to perform a little better then FS1 and FS3, especially under the dominant model. This seems the only pattern that is consistent across most simulations. For more examples refer to appendix C. In conclusion there is no clear-cut pattern that would prefer a certain family structure. In practice this allows to recruit the most convenient family structure as suggested by other aspects. Also mixtures of family structures can be sampled. One example is the Alzheimer's data set used in this thesis (section 9, page 87).

8.5. **Haplotype analysis.** Tables C.30 - C.32 (pp. 168 - 173) display haplotype simulations of random samples. Simulations for ascertained samples are shown in tables C.33 - C.35 (pp. 175 - 179). These simulations show a similar pattern to the simulations already described. It is noteworthy that the allele frequency of the disease associated allele is estimated very accurately. This reflects the fact that every haplotype contributes to the estimation of this parameter. Intrestingly, this holds even for misspecification with the exception of data simulated under the recessive model. In this instance the allele frequency p^* is moderately overestimated in the cases considered (for example table 8.4). In the haplotype analysis it becomes more apparent that under misspecification all parameters except for β, p^* have almost unbiased estimates. The pattern of biases for β and p^* as seen for single locus analysis in tables C.2 - C.19 can be summarized as follows (with the following symbols denoting relative assessments: $\sim$: roughly unbiased, $\nearrow$: weak bias upwards, $\uparrow$: stronger bias upwards, $\searrow$: weak bias downwards, $\downarrow$: stronger bias downwards):

Simulation	**Estimation**	p^*	β
dominant	recessive	$\sim$	$\uparrow$
dominant	additive	$\searrow$	$\nearrow$
recessive	dominant	$\uparrow$	$\uparrow$
recessive	additive	$\uparrow$	$\uparrow$
additive	dominant	$\sim$	$\downarrow$
additive	recessive	$\sim$	$\downarrow$

These patterns can help to assess possible errors in practical data analysis. In conclusion, under the alternative, considering haplotypes reduces MSE for β and p^* by tendency as compared to single locus analysis.

8.6. **Asymptotic normality of parameter estimates.** To evaluate the validity of confidence intervals constructed from the model-based Fisher information (section 5.4.1) the model based Fisher information was compared to empirical Fisher information. Both are asymptotically identical in probability. The speed of convergence can therefore be assessed by this comparison. Model based estimates are either derived from the first

Table 8.4: Parameter estimates for a total of N = 1920 individuals and random sampling. MSE is given in parentheses in a separate line. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$\eta_2 = 0.2$	$\eta_3 = 0.3$	$R_1 = 0.1$	$R_2 = 0.2$	$R_3 = -0.0$	$p^\star = 0.3$	$\beta = 2.0$
simulation: dom; estimation: rec									
FS1a	960	0.22	0.15	0.29	0.15	0.37	-0.03	0.31	0.21
		(0.00)	(0.00)	(0.00)	(0.04)	(0.09)	(0.03)	(0.00)	(3.24)
FS3a	720	0.22	0.15	0.29	0.15	0.35	-0.03	0.31	0.66
		(0.00)	(0.00)	(0.00)	(0.04)	(0.05)	(0.03)	(0.00)	(1.82)
FS4a	720	0.22	0.15	0.29	0.14	0.42	-0.04	0.30	0.32
		(0.00)	(0.00)	(0.00)	(0.04)	(0.08)	(0.03)	(0.00)	(2.83)
FS6a	576	0.22	0.15	0.29	0.14	0.34	-0.04	0.30	0.68
		(0.00)	(0.00)	(0.00)	(0.04)	(0.05)	(0.03)	(0.00)	(1.77)
simulation: rec; estimation: rec									
FS1a	960	0.22	0.15	0.29	0.18	0.39	-0.08	0.29	1.01
		(0.00)	(0.00)	(0.00)	(0.05)	(0.07)	(0.04)	(0.00)	(1.35)
FS3a	720	0.22	0.15	0.29	0.10	0.24	-0.02	0.34	1.93
		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)	(0.01)	(0.48)
FS4a	720	0.22	0.15	0.29	0.15	0.35	-0.07	0.28	1.39
		(0.00)	(0.00)	(0.00)	(0.04)	(0.05)	(0.03)	(0.01)	(0.80)
FS6a	576	0.22	0.15	0.29	0.10	0.24	-0.03	0.34	1.90
		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.01)	(0.01)	(0.46)
simulation: add; estimation: rec									
FS1a	960	0.22	0.15	0.29	0.14	0.41	-0.04	0.31	0.19
		(0.00)	(0.00)	(0.00)	(0.04)	(0.10)	(0.02)	(0.00)	(3.28)
FS3a	720	0.22	0.15	0.29	0.13	0.37	-0.03	0.32	0.53
		(0.00)	(0.00)	(0.00)	(0.03)	(0.05)	(0.02)	(0.00)	(2.17)
FS4a	720	0.22	0.15	0.29	0.14	0.44	-0.04	0.31	0.28
		(0.00)	(0.00)	(0.00)	(0.04)	(0.09)	(0.03)	(0.00)	(2.96)
FS6a	576	0.22	0.15	0.29	0.13	0.36	-0.04	0.32	0.54
		(0.00)	(0.00)	(0.00)	(0.03)	(0.05)	(0.02)	(0.00)	(2.15)

$(\hat{\mathrm{I}}_1)$ or second derivatives $(\hat{\mathrm{I}}_2)$ of the log likelihood. For log likelihood l, MLE $\hat{\theta}$ and observation $X^{(1)}$ we have: $\hat{\mathrm{I}}_1 = \frac{\partial}{\partial\theta_1} l(\theta; X^{(1)})(\frac{\partial}{\partial\theta} l(\theta; X^{(1)}))^T|\hat{\theta}_1$ and $\hat{\mathrm{I}}_2 = -\frac{\partial^2}{\partial\theta^2} l(\theta; X^{(1)})|\hat{\theta}_1$. For M independent replications $X^{(1)}, ..., X^{(M)}$ the mean is taken to estimate the Fisher information of the likelihood in θ. The empirical Fisher information is derived from the parameter estimates and does not refer to the likelihood: $\hat{\mathrm{I}}_{emp} = \left\{M\hat{\mathrm{Cov}}(\hat{\theta})\right\}^{-1}$, $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_M)$, where $\hat{\mathrm{Cov}}(\hat{\theta})$ is the sample covariance matrix of $\hat{\theta}$. For likelihood (5.8) and data replicates $(Y^{(j)}, G^{(j)}), j = 1, ..., M$, families $i = 1, ..., N$ we have (numeric algorithm in appendix C.2):

$$\begin{aligned}
\hat{\mathrm{I}}_1 &= \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\left\{(\frac{\partial}{\partial\theta}l(\theta|Y_i^{(j)}, G_i^{(j)}))(\frac{\partial}{\partial\theta}l(\theta|Y_i^{(j)}, G_i^{(j)}))^T|\hat{\theta}_j\right\},\\
\hat{\mathrm{I}}_2 &= -\frac{1}{M}\sum_{j=1}^{M}\frac{1}{N}\frac{\partial^2}{\partial\theta^2}l(\hat{\theta_m}|Y^{(j)}, G^{(j)}) = -\frac{1}{M}\sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\partial^2}{\partial\theta^2}l(\theta|Y_i^{(j)}, G_i^{(j)})|\hat{\theta}_j\right\},
\end{aligned}$$

where $l(\theta|Y_i^{(j)}, G_i^{(j)})$ is the likelihood contribution of family i in replication j. Evaluation takes place at $\hat{\theta}_j$

Table 8.5 shows estimates for simulations with $N = 10.000$ families and $M = 500$ replications. Entry-wise ratios $\hat{\mathrm{I}}_1/\hat{\mathrm{I}}_{emp}$ and $\hat{\mathrm{I}}_2/\hat{\mathrm{I}}_{emp}$ are displayed in table 8.6. The diagonal elements show close agreement. Since agreement is worse for sample sizes < 10.000, normality of parameter estimates should not be assumed for small sample sizes. Instead bootstrap estimates should be used to construct confidence intervals. Table 8.7 shows the empirical covariances as well as estimates based on the two estimates of the Fisher-information together with the robust sandwich estimator $\hat{\mathrm{I}}_S := \hat{\mathrm{I}}_2^{-1}\hat{\mathrm{I}}_1(\hat{\mathrm{I}}_2^{-1})^T$. The sandwich estimator is known to yield consistent estimates for heteroscedastic as well as for certain classes of dependant data [42]. It is therefore worthwhile to consider the sandwich estimator in situations where distributional assumptions are violated, *e.g.* lack of convergence. However, the sandwich estimator behaves very similar to the model based estimators and can therefore not help to construct reliable confidence intervals.

Table 8.5: Estimates of the Fisher information for a random sample with $N = 10.000$ families with two offspring. Baseline parameter $\mu = -2$. Model of inheritance is additive. The empirical Fisher information is based on $M = 500$ replications.

	$\eta = 0.3$	$R = 0.5$	$p^\star = 0.3$	$\beta = 1$
	$\hat{\mathrm{I}}_1$			
η	19.8	-0.086	-1.39	-0.539
δ	-0.086	0.28	0.036	0.114
$p^\star$	-1.39	0.036	2.79	1.017
β	-0.539	0.114	1.017	0.438
	$\hat{\mathrm{I}}_2$			
η	19.8	-0.085	-1.39	-0.536
δ	-0.085	0.28	0.036	0.114
$p^\star$	-1.39	0.036	2.79	1.016
β	-0.536	0.114	1.016	0.437
	$\hat{\mathrm{I}}_{emp}$			
η	20.3	0.05	-1.23	-0.44
δ	0.05	0.25	-0.053	0.072
$p^\star$	-1.23	-0.053	2.57	0.92
β	-0.44	0.072	0.92	0.382

Table 8.6: Ratios of estimated Fisher matrices. Estimates of the Fisher information for a random sample with $N = 10.000$ families with two offspring. Baseline parameter $\mu = -2$. Model of inheritance is additive. The empirical Fisher information is based on $M = 500$ replications.

	$\eta = 0.3$	$R = 0.5$	$p^\star = 0.3$	$\beta = 1$
	$\hat{\mathrm{I}}_1/\hat{\mathrm{I}}_{emp}$			
η	0.97	-1.72	1.13	1.22
δ	-1.72	1.12	-0.67	1.57
$p^\star$	1.13	-0.67	1.08	1.1
β	1.22	1.57	1.1	1.14
	$\hat{\mathrm{I}}_2/\hat{\mathrm{I}}_{emp}$			
η	0.97	-1.69	1.12	1.22
δ	-1.69	1.12	-0.68	1.58
$p^\star$	1.12	-0.68	1.08	1.1
β	1.22	1.58	1.1	1.14

Table 8.7: Estimates of the covariance matrices.Estimates of the Fisher information for a random sample with $N = 10.000$ families with two offspring. Baseline parameter $\mu = -2$. Model of inheritance is additive. The empirical Fisher information is based on $M = 500$ replications.

	$\eta = 0.3$	$R = 0.5$	$p^\star = 0.3$	$\beta = 1$
	$\mathbf{Cov}_{emp}$			
η	0.05	-0.01	0.01	0.02
δ	-0.01	10.43	6.51	-17.62
$p^\star$	0.01	6.51	6.83	-17.62
β	0.02	-17.62	-17.62	48.3
	$\hat{\mathrm{I}}_1^{-1}$			
η	0.05	0.01	0.02	0.01
δ	0.01	7.78	4.19	-11.74
$p^\star$	0.02	4.19	4.62	-11.8
β	0.01	-11.74	-11.8	32.78
	$\hat{\mathrm{I}}_2^{-1}$			
η	0.05	0.01	0.02	0.01
δ	0.01	7.77	4.18	-11.74
$p^\star$	0.02	4.18	4.61	-11.78
β	0.01	-11.74	-11.78	32.72
	$\left\{\hat{\mathrm{I}}_2^{-1}\hat{\mathrm{I}}_1(\hat{\mathrm{I}}_2^{-1})^T\right\}^{-1}$			
η	0.05	0.01	0.02	0.01
δ	0.01	7.77	4.18	-11.74
$p^\star$	0.02	4.18	4.60	-11.73
β	0.01	-11.73	-11.75	32.67

8.7. **Power comparison with the FBAT statistic.** The LR-Test (sec. 5.5) was compared to a robust test statistic. This allows to judge effects on power and significance level maintenance of the LR-test relative to the robust method. In order to achieve an accurate comparison, the tests were applied to the same data sets. The estimator for parameter R_1 was already shown to be unbiased under misspecification of modes of inheritance under the null hypothesis earlier. Therefore the significance level should be accurately maintained. The FBAT statistic is robust both to misspecification of modes of inheritance and population stratification. Recall, that FBAT scores transmissions of alleles/haplotypes from parents to offspring similar to the TDT (section 3.2.2). For the comparisons, a single observed locus was simulated in order to allow a direct comparison by avoiding the correction for multiple testing. Population stratification is not considered here and is discussed later. Misspecification always refers to the mode of inheritance in the remainder of this section. The significance level was $\alpha = 0.05$ in all simulations. If not otherwise stated, both parents were simulated in the nuclear families.

Table 8.8: Simulations to evaluate maintenance of significance level for LR-Test and FBAT (see text). $n = 400$ nuclear families with one offspring were simulated, $\mu = -2.00$, the number of iterations is $N = 5000$. *sim* and *est* denote the modes of inheritance under which the data was generated and the test was performed, respectively. For the ascertained case (columns LR^a and $FBAT^a$) two offspring were simulated.

	η_1	R_1	$p^\star$	β	sim	est	LR	FBAT	LR^a	FBAT^a
1	0.3	0.0	0.3	2.0	dom	dom	0.053	0.047	0.058	0.060
2	0.3	0.0	0.3	2.0	dom	rec	0.051	0.047	0.053	0.060
3	0.3	0.0	0.3	2.0	dom	add	0.050	0.047	0.053	0.060
4	0.3	0.0	0.3	2.0	rec	dom	0.067	0.051	0.076	0.054
5	0.3	0.0	0.3	2.0	rec	rec	0.051	0.051	0.051	0.054
6	0.3	0.0	0.3	2.0	rec	add	0.059	0.051	0.061	0.054
7	0.3	0.0	0.3	2.0	add	dom	0.051	0.051	0.051	0.064
8	0.3	0.0	0.3	2.0	add	rec	0.053	0.051	0.053	0.064
9	0.3	0.0	0.3	2.0	add	add	0.052	0.051	0.049	0.059
10	0.2	0.0	0.2	2.0	dom	dom	0.054	0.045	0.058	0.056
11	0.2	0.0	0.2	2.0	dom	rec	0.153	0.045	0.057	0.056
12	0.2	0.0	0.2	2.0	dom	add	0.054	0.045	0.051	0.056
13	0.2	0.0	0.2	2.0	rec	dom	0.035	0.050	0.057	0.054
14	0.2	0.0	0.2	2.0	rec	rec	0.048	0.050	0.066	0.054

15	0.2	0.0	0.2	2.0	rec	add	0.036	0.050	0.055	0.054
16	0.2	0.0	0.2	2.0	add	dom	0.053	0.046	0.055	0.057
17	0.2	0.0	0.2	2.0	add	rec	0.131	0.046	0.061	0.057
18	0.2	0.0	0.2	2.0	add	add	0.055	0.046	0.056	0.057

8.7.1. *Maintenance of significance level under the null hypothesis.* Table 8.8 shows result for simulations under the null hypothesis. Both tests maintain the significance level both under correct specification and misspecification, for high allele frequencies (simulations 1-9). For lower allele frequencies (simulations 10-18), with the exception of two recessive model fits of the LR statistic (simulations 11 and 17). These exceptions are due to numerical instabilities, since random sampling provides very few affecteds for low allele frequencies making the numerical optimization difficult. It should be noted that these two simulations do not represent practically relevant situations. In ascertained samples all simulations show accord with the α level.

8.7.2. *Power comparisons.* Power comparisons are shown in table 8.9. In all cases that were considered the LR-test performed better than FBAT. Although this result is expected since FBAT conditions on phenotypes and parental genotypes, the differences are quite remarkable. For example, in simulation 10 (dominant model, $R_1 = 0.2$) the power of the LR test is 72% and compares to 18% of FBAT (random) or 81% as compared to 37% (ascertained).

Without ascertainment (columns LR and FBAT) both statistics seem to be quite insensitive to model misspecification. Under ascertained sampling (columns LR^a and FBAT^a) FBAT again is almost insensitive to misspecification, whereas the LR test shows a slight decrease in power under misspecification under the recessive model.

For several cases a different family structure was chosen and compared in columns LR^{a1} and FBAT^{a1}. For these columns, families with one parent and three offspring were ascertained on one affected offspring (as compared with two parents and two offspring for columns LR^a and FBAT^a). Again, power is much better for the LR test, although the difference is not as pronounced as in the cases considered so far (simulations 1, 2). In

Table 8.9: For a description of the columns see caption of table 8.8. $n = 400$, $\mu = -2.00$ and $N = 5000$ as under the null hypothesis. Columns LR and FBAT list simulations for random samples as in table 8.8, columns LR^a and FBAT^a list ascertained samples with two offspring and columns LR^{a1} and FBAT^{a1} enumerate simulations with one parent and 3 offspring.

	η_1	R_1	$p^\star$	β	sim	est	LR	FBAT	LR^a	FBAT^a	LR^{a1}	FBAT^{a1}
1	0.3	0.1	0.3	2.0	dom	dom	0.251	0.082	0.304	0.134	0.265	0.172
2	0.3	0.1	0.3	2.0	dom	rec	0.234	0.082	0.229	0.129	0.240	0.172
3	0.3	0.1	0.3	2.0	dom	add	0.240	0.080	0.274	0.129	-	-
4	0.3	0.1	0.3	2.0	rec	dom	0.150	0.060	0.201	0.099	-	-
5	0.3	0.1	0.3	2.0	rec	rec	0.122	0.060	0.198	0.099	-	-
6	0.3	0.1	0.3	2.0	rec	add	0.136	0.060	0.203	0.099	-	-
7	0.3	0.1	0.3	2.0	add	dom	0.401	0.118	0.510	0.221	0.423	0.194
8	0.3	0.1	0.3	2.0	add	rec	0.376	0.118	0.446	0.221	0.439	0.294
9	0.3	0.1	0.3	2.0	add	add	0.404	0.118	0.524	0.221	-	-
10	0.3	0.2	0.3	2.0	dom	dom	0.719	0.176	0.812	0.373	0.722	0.479
11	0.3	0.2	0.3	2.0	dom	rec	0.669	0.181	0.691	0.373	-	-
12	0.3	0.2	0.3	2.0	dom	add	0.705	0.176	0.776	0.373	0.715	0.478
13	0.3	0.2	0.3	2.0	rec	dom	0.302	0.092	0.520	0.239	0.421	0.298
14	0.3	0.2	0.3	2.0	rec	rec	0.305	0.094	0.586	0.239	0.524	0.274
15	0.3	0.2	0.3	2.0	rec	add	0.316	0.094	0.545	0.239	0.412	0.276
16	0.3	0.2	0.3	2.0	add	dom	0.915	0.303	0.972	0.646	0.929	0.753
17	0.3	0.2	0.3	2.0	add	rec	0.897	0.295	0.954	0.646	0.941	0.684
18	0.3	0.2	0.3	2.0	add	add	0.923	0.295	0.976	0.646	-	-

the cases considered here FBAT shows more variability than the LR test (simulations 7-8, 16-17). It is interesting to see that the robust statistic is more sensitive to misspecification than the parametric LR test in terms of power. This is not further explored in this thesis but deserves further consideration.

In conclusion the power of the LR test is better in all cases considered (factor of $\sim 2-4$). The LR test shows comparable robustness to misspecification as the FBAT statistic, as far as power is concerned. This is due to the fact that the LR statistic uses information about the marginal parental genotype distribution and parental phenotypes. Therefore, if population stratification is unlikely, the LR test is preferable.

8.8. **Random effects model.** The one-dimensional random effects model established in section 5.3.2 was not simulated to estimate MSEs because of the computational burden.

Table 8.10: Parameter fits for the random effects model for several data sets. True parameters in parenthesis. n families are simulated with 4 offspring. The ascertainment event is on one affected offspring. Baseline $\mu = -2$.

n	η_1	R_1	$p^\star$	β	σ
1000	0.40 (0.4)	-0.05 (0.0)	0.47 (0.4)	2.7 (3.0)	1.05 (1.2)
1000	0.41 (0.4)	0.01 (0.0)	0.50 (0.4)	2.5 (3.0)	0.90 (1.2)
1000	0.38 (0.4)	-0.01 (0.0)	0.45 (0.4)	2.8 (3.0)	1.15 (1.2)
1000	0.30 (0.3)	0.52 (0.5)	0.30 (0.3)	3.0 (3.0)	1.26 (1.2)
1000	0.29 (0.3)	0.50 (0.5)	0.32 (0.3)	3.1 (3.0)	0.76 (1.2)
1000	0.31 (0.3)	0.50 (0.5)	0.31 (0.3)	2.9 (3.0)	1.17 (1.2)
1000	0.30 (0.3)	0.49 (0.5)	0.27 (0.3)	3.1 (3.0)	0.01 (0.2)
1000	0.29 (0.3)	0.50 (0.5)	0.31 (0.3)	2.9 (3.0)	0.18 (0.2)
1000	0.30 (0.3)	0.50 (0.5)	0.34 (0.3)	2.8 (3.0)	0.02 (0.2)

Rather parameter fits are shown under the null-hypothesis and the alternative to make identifiability plausible. The Alzheimer's data set is later analyzed using this model. Table 8.10 lists some parameter fits. From this table it seems that the random effect is more difficult to estimate than the other parameters. Especially for a small random effect (0.2), the effect can be underestimated.

8.9. **MCMC simulations.** The MCMC framework with collapsed haplotypes offers an elegant means of estimating association parameters. Due to the collapsing step, only random samples can be handled by this approach. The algorithm is easy to implement. However, it is computationally very demanding. Plots of an example run are given in figure 8.10. In this instance the chain quickly finds the correct range of the parameters. The rejection rates for diplotypes and β were 27.9% and 29.1%, respectively. The prior distributions for haplotype frequencies were chosen as being almost uninformative: $\eta_i^\star \sim$ Dirichlet$(1.5, ..., 1.5)$. For β a mild prior was chosen for β: $\beta \sim N(1.5, 3)$. Results of simulations for random samples are shown in table 8.11. Exemplary convergence plots are shown in the appendix (figures C.1 - C.3, pp. 182 - 184). The simulations are very accurate under the correct model (first half of the table). However, for the misspecified model with simulation of recessive inheritance and estimation of additive inheritance, β is strongly biased. The bias exceeds that of a comparable simulation for the frequentist's

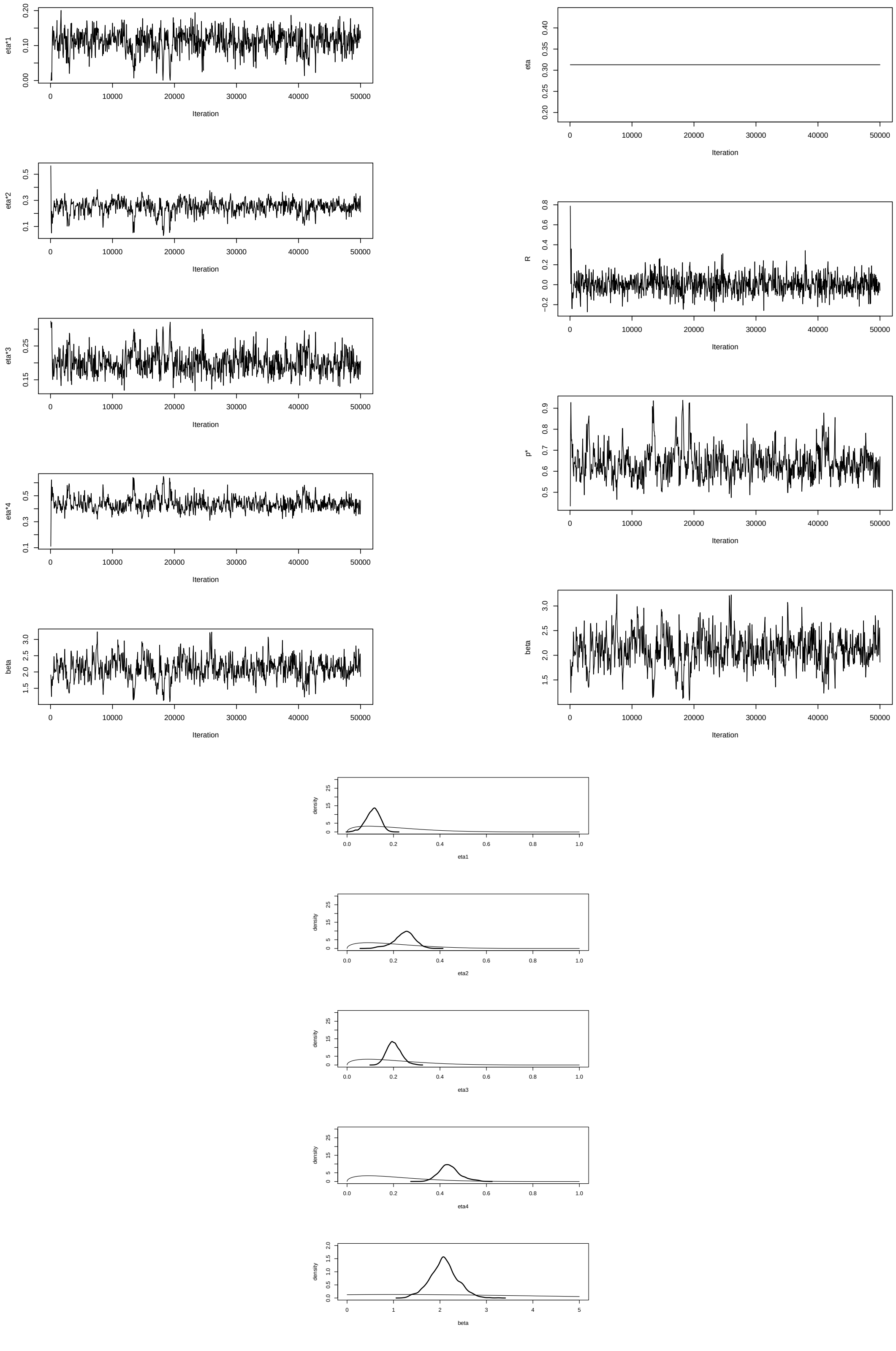
Table 8.11: Bayesian simulations for n randomly sampled families. Baseline $\mu = -2$. Burn in 25000 iterations, estimation 25000 iterations. True parameter values are given in parenthesis. 95% credible intervals of the estimated posterior distribution are given in an extra line.

-	n	$\eta_1^\star = 0.1$	$\eta_2^\star = 0.7$	$\eta_3^\star = 0.1$	$\eta_4^\star = 0.1$	$\beta = 2.0$
	simulation: add; estimation: add					
FS3	400	0.10 (0.09)	0.71 (0.70)	0.10 (0.11)	0.08 (0.09)	2.04 (2.00)
		(0.00, 0.15)	(0.60, 0.76)	(0.05, 0.20)	(0.04, 0.19)	(0.99, 3.20)
FS4	400	0.05 (0.09)	0.64 (0.72)	0.18 (0.11)	0.13 (0.08)	1.59 (2.00)
		(0.00, 0.12)	(0.54, 0.70)	(0.11, 0.24)	(0.07, 0.23)	(1.08, 2.29)
FS5	400	0.04 (0.08)	0.60 (0.74)	0.24 (0.10)	0.12 (0.08)	1.42 (2.00)
		(0.00, 0.13)	(0.52, 0.67)	(0.15, 0.29)	(0.06, 0.19)	(1.06, 2.14)
FS6	300	0.11 (0.09)	0.74 (0.73)	0.09 (0.10)	0.06 (0.08)	2.44 (2.00)
		(0.07, 0.13)	(0.70, 0.76)	(0.06, 0.12)	(0.04, 0.10)	(1.77, 3.09)
	simulation: rec; estimation: add					
FS3	400	0.06 (0.09)	0.55 (0.70)	0.14 (0.11)	0.24 (0.09)	15.52 (2.00)
		(0.04, 0.08)	(0.52, 0.59)	(0.12, 0.16)	(0.21, 0.28)	(3.72, 31.23)
FS4	400	0.01 (0.09)	0.43 (0.72)	0.22 (0.11)	0.34 (0.08)	2.72 (2.00)
		(0.00, 0.04)	(0.35, 0.50)	(0.19, 0.24)	(0.27, 0.41)	(2.12, 3.81)
FS5	400	0.00 (0.08)	0.40 (0.74)	0.28 (0.10)	0.32 (0.08)	2.45 (2.00)
		(0.00, 0.02)	(0.34, 0.46)	(0.25, 0.30)	(0.26, 0.38)	(2.02, 2.93)
FS6	300	0.02 (0.09)	0.48 (0.73)	0.18 (0.10)	0.32 (0.08)	2.95 (2.00)
		(0.00, 0.06)	(0.40, 0.55)	(0.14, 0.20)	(0.25, 0.40)	(2.08, 4.79)

model and therefore indicates convergence problems of the chain. Figure C.3 (page 184) indicates, that the main problem lies with β and that the chain might eventually converge. Speeding up convergence is subject to further research and is discussed later.

8.10. **Computational issues.** One reason for the large variance in the posterior of β in case of misspecification (recessive/additive) seems to be that rejection rates, both, for β and H^* are below 5% and convergence of the chain is not achieved in the 50,000 iterations that were employed in the simulations. The proposal distribution of β is $N(\beta_O, .1)$ with β_O being the old value of β. An attempt was made to improve convergence by increasing the variance of the proposal distribution for β. A variance of .5 already yields a rejection rate of $\sim$ 70% for β and a variance of .75 corresponds to a rejection rate of 80%. However, the rejection rate for H^* is still below 5% in all cases that were considered. It is therefore necessary to employ more sophisticated changes to the sampler which are discussed later. I note, that the completed data of the collapsed sampler allows for an elegant computation of the acceptance probability $P(Y, H^\star; \beta_N)$ which can be factorized into $P(Y; \beta_N|H^\star)P(H^\star)$; the haplotype part cancels out in the Metropolis step. The phenotype part (term A in (7.4), page 70) can be computed efficiently.

Figure 8.2: Display of the markov chain of the MCMC sampler. Top left is the original chain $(\eta_1^*, \eta_2^*, \eta_3^*, \eta_4^*, \beta) = (.15, .25, .2, .4, 2)$. Top right shows the reparametrized chain in terms of parametrization 4.The bottom diplays the prior (thin line) and a density estimate of the posterior distributions (thick line).



9. Alzheimer's disease

Alzheimer's disease is a neuro-degenerative disorder that has been described by Alois Alzheimer in 1906 and was characterized to be an early onset dementia accompanied by typical changes of the brain: the loss of neuron mass (atrophy) and the formation of protein aggregates (plaques). Today Alzheimer's is defined by the presence of dementia and characteristic brain changes. Alzheimer cases are classified to be early or late onset by onset of earlier or later than at the age of 65, respectively. By the frequency of the disease (prevalence 0.5%) it is a complex disorder by definition, but it also shows familial aggregation. This allowed the definition of eight loci implicated with Alzheimer's labeled AD1 through AD8 [69, 70]. Mutations have been identified in four genes (AD1-AD4), however, four other loci are implicated with Alzheimer's disease (table 9.1).

Table 9.1: Known genetic loci linked with Alzheimer's disease

Name	**MIM**	**Locus**	**Gene**
AD1	104760	21q21	amyloid precursor gene (APP)
AD2	107741	19q13.2	apolipoprotein E (ApoE)
AD3	607822	14q24.3	presenilin-1 (PSEN1)
AD4	600759	1q31-q42	presenilin-2 (PSEN2)
AD5	602096	12p11.23-q13.12	-
AD6	605526	10q24	-
AD7	606187	10p13	-
AD8	607116	20p12.2-q11.21	-

Thus Alzheimer's is a heterogeneous disorder, *i.e.* several genes give rise to an identical phenotype without apparent interaction. 25% of all Alzheimer cases involve familial aggregation. Therefore 75% are most likely due to genes with smaller effects that prevent family clusters from forming. In the following, we focus on the ApoE locus, since the available data set contains information about this genomic region.

9.1. **The ApoE locus.**

AD2 is a late onset form of Alzheimer's disease, implying that the age of onset of most cases is greater than 65 years of age. Linkage of AD2 to chromosome 19 was first shown in 1989 and later the ApoE variation was characterized to be associated with Alzheimer's disease [76, 15]. The ApoE protein has been characterized as being

Table 9.2: Characteristics of the ApoE/Alzheimer's data set. *n (total)* is the number of individuals in the data set. Missing phenotypes are calculated for a cut-off age of 50.

Parameter	**Value**
n (families)	131
n (total)	656
Missing Genotypes	46.5%
Missing Phenotypes	5.3%

important in lipid metabolism. Lipid composites contained in the blood harbor proteins that serve as signals for lipid uptake into cells. Among them is ApoE and two receptors have been characterized that bind ApoE: the low density lipoprotein (LDL) receptor and an ApoE specific receptor [19]. On account of this knowledge a sizable amount of research has been invested into relationships of ApoE and cardiovascular disease [19]. The gene harbors three major alleles called $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$. These alleles are formed from haplotypes that are composed of SNPs which induce substitutions at positions 112 and 158 in the resulting amino acid sequence. Association studies for Alzheimer's disease have found a dosage effect of the $\epsilon 4$ allele, *i.e.* the penetrance function is of the form $0 < f(0) < f(1) < f(2) < 1$, where the argument is the number of $\epsilon 4$ alleles in the genotype and $f(i)$ specifies the probability of being affected of Alzheimer's at a certain age.

9.2. **The ApoE/Alzheimer's data set.** The data set used in this thesis, was first published in a study scrutinizing the effect of typing SNPs that are in LD with a disease SNP (or haplotype) [64] which is similar to the focus of this thesis. A subset, both, in terms of markers and probands has been made available for applying the methods in this thesis. The characteristics of the ApoE/Alzheimer's data set are shown in table 9.2. It is noteworthy that a substantial amount of data is missing. Missing parental genotypes account for most of the missing data. For non-deceased persons, disease status can not be readily assessed on account of the late onset of Alzheimer's disease. The approach taken here was to choose a cut-off age above which proband were defined to be unaffected and below which the phenotype was treated as missing. The cutoff value used in the analysis

Figure 9.1: Distribution of number of offspring in the ApoE/Alzheimer's data set. Family size includes parents.

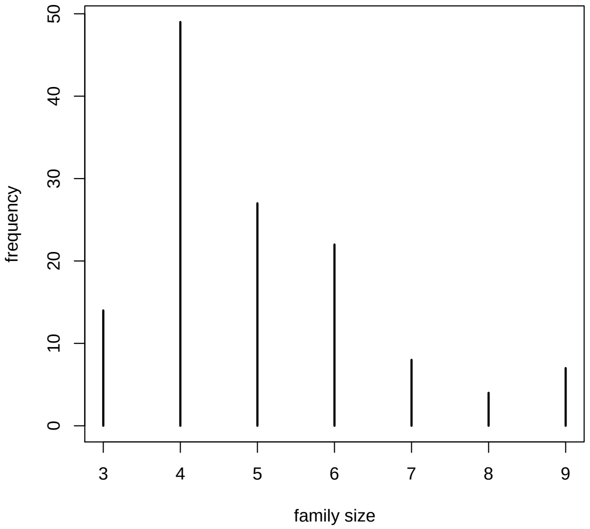
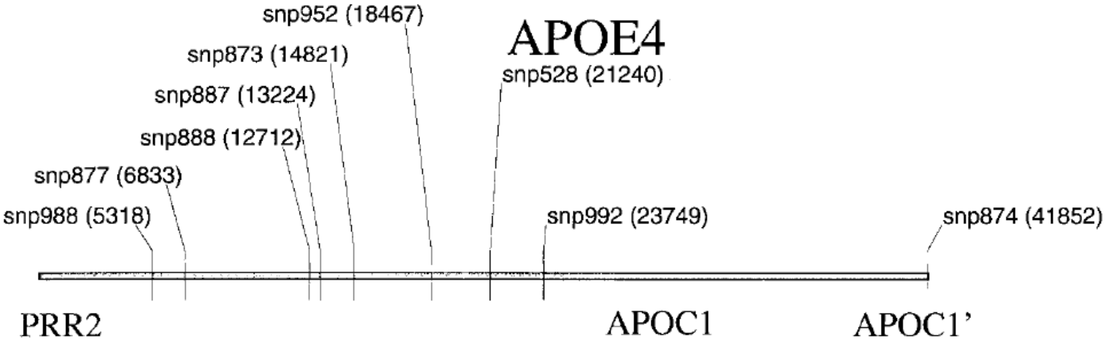


Figure 9.2: Physical map of the neighborhood of the ApoE locus. Base pairs are given in parenthesis relative to the origin of sequencing.



was 60 years of age. Missing data is treated by marginalizing the likelihood over missing data. Family size varies considerably in families and the distribution is shown in figure 9.2.

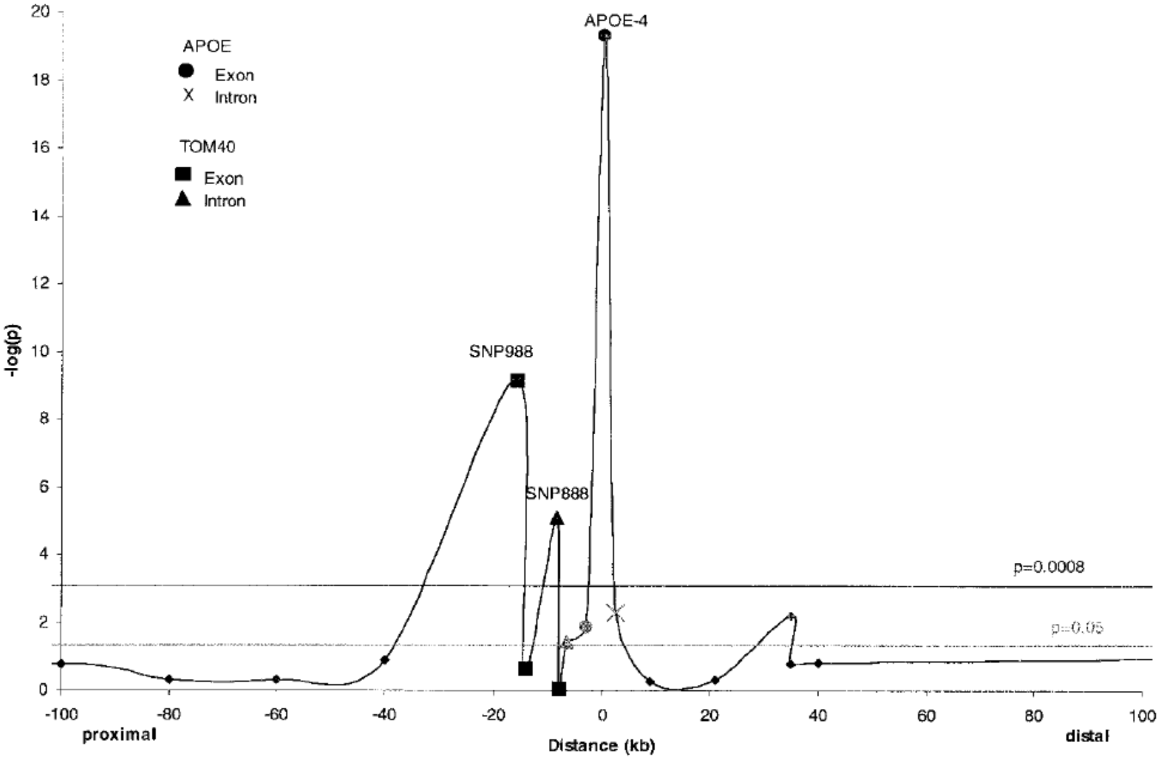
The physical ordering of some of the SNPs is displayed in figure 9.2. SNPs that were used in this study, given in physical ordering, are *SNP1006, SNP875, SNP886, SNP988,* *SNP888, SNP873, SNP952, SNP528, SNP992, SNP465, SNP457, SNP471, SNP479, SNP497, SNP491, SNP459, SNP512.* SNP $SNP528$ is located within the ApoE gene. The distance of SNP $SNP512$ and $SNP1006$ from SNP $SNP528$ is about 1,520 kb and 100 kb, respectively.

The data set was fitted to additive, dominant and recessive modes of inheritance. Results for the single locus analysis are shown in table 9.3. For testing, a likelihood ratio test was performed with the null hypothesis $R_1 = ... = R_M = 0$, where $M = 2^K$ is the number of correlation parameters.

9.2.1. *Single locus analysis.* Since the recessive model did not provide a good fit in most instances, results are reported in the appendix (table D.1, page 185). Like the recessive model, the additive model results in parameter estimates on the border of the parameter space. The dominant model, however, shows plausible parameter estimates in all cases. As the dominant mode of inheritance shows the best fit in terms of likelihood values in MLEs (table 9.3), this model is favored for data interpretation.

Figure 9.3 (page 91) shows results of a case-control study of Alzheimer's disease [64]. The strongest associations in this study were seen with $SNP528$, $SNP988$ and $SNP888$. This matches with the findings for the dominant model reported in table 9.3. For $SNP528$ the parameter R_1 is 0.9, supporting the hypothesis that $SNP528$ is located within the same gene as the disease variant. Since the SNPs composing the ϵ-Alleles are not contained in the data set, it is not possible to directly compare estimates with those for the variants believed to be causative. The effect sizes estimated for the dominant model vary between 1.24 and 1.46 (corresponding to odds ratios (ORs) between 3.46 and 4.31). These estimates are therefore consistent across different loci. Estimates of $p^\star$ vary between 0.43 and 0.82 showing a greater variability. Since $SNP528$ is strongly correlated with the marker represented by $p^\star$, the estimates of $p^\star$ and β derived from $SNP528$ should be the most accurate, placing $p^\star$ at 0.43 and β at 1.46. The random effects model (5.3.2; page 45) was fitted to the data set. Results are shown in table 9.5. The random effect is generally estimated to be very small (< 0.03). This implies that the effect of the ApoE locus

Figure 9.3: Results of a case control study in Alzheimer's disease in the ApoE region. The x-axis shows location of SNPs with respect to the ApoE gene and the y-axis shows negative logarithms of corresponding p-values.



seems to be large as compared to other loci and thereby parameter estimates for the other parameters are almost unchanged as compared to the model without random effect. It should be noted that the low number of families (n=131) makes the random effect hard to estimate.

Confidence intervals (CIs) for $SNP528$ were based on quantiles of the empirical bootstrap distribution function using 1000 bootstrap repetitions and were $\hat{\eta}_1 = 0.38$, (95%-CI 0.32, 0.44), $\hat{R}_1 = 0.89$, (95%-CI 0.71, 0.99), $\hat{p}^* = 0.43$, (95%-CI 0.34, .53) and $\hat{\beta} = 1.46$, (95%-CI 1.29, 1.66). A fit of the random effects model for $SNP528$ gave the estimates $(\hat{\eta}_1, \hat{R}_1, \hat{p}^*, \hat{\sigma}_a) = (0.38, 0.89, 0.43, 1.47, 0.007)$.

Table 9.3: Data analysis for a single SNP analysis of the Alzheimer's data set. l_A is the supremum of the likelihood under the alternative; the p-value is based on a χ^2_1 approximation of the LR-statistics. Parameter estimates are given under the alternative. Significant p-values are given in bold.

	SNP	model	l_A	p-value	η_1	R_1	$p^\star$	β
1	SNP1006	add	-6.734e+02	9.9e-01	0.66	0.00	1.00	0.59
2		dom	-6.717e+02	7.9e-02	0.64	0.82	0.72	1.25
4	SNP875	add	-6.678e+02	9.9e-01	0.70	0.00	1.00	0.59
5		dom	-6.674e+02	4.6e-01	0.70	-0.36	0.76	1.24
7	SNP886	add	-6.588e+02	9.9e-01	0.71	0.00	1.00	0.59
8		dom	-6.580e+02	2.6e-01	0.72	-0.43	0.68	1.28
10	SNP988	add	-6.781e+02	**1.0e-02**	0.59	-0.65	0.62	0.81
11		dom	-6.735e+02	**8.7-05**	0.59	-0.85	0.49	1.41
13	SNP888	add	-6.637e+02	**2.8e-02**	0.30	-0.87	0.76	0.70
14		dom	-6.625e+02	**9.2e-03**	0.28	-0.64	0.57	1.34
16	SNP873	add	-6.549e+02	**1.9e-02**	0.73	0.74	0.83	0.69
17		dom	-6.545e+02	**1.4e-02**	0.73	0.59	0.56	1.35
19	SNP952	add	-6.692e+02	8.2e-01	0.72	0.18	0.99	0.60
20		dom	-6.684e+02	2.4e-01	0.71	0.49	0.68	1.27
22	SNP528	add	-6.837e+02	**8.0e-04**	0.38	0.72	0.54	0.89
23		dom	-6.766e+02	**6.5e-07**	0.38	0.89	0.43	1.47
25	SNP992	add	-6.464e+02	9.9e-01	0.27	-0.00	1.00	0.59
26		dom	-6.447e+02	6.9e-02	0.29	-0.87	0.76	1.24
28	SNP465	add	-7.072e+02	9.9e-01	0.44	-0.00	1.00	0.59
29		dom	-7.069e+02	7.9e-01	0.44	-0.12	0.72	1.25
31	SNP457	add	-7.202e+02	8.5e-01	0.51	-0.13	0.98	0.60
32		dom	-7.197e+02	4.1e-01	0.52	-0.25	0.64	1.30
34	SNP471	add	-7.171e+02	9.9e-01	0.54	0.00	1.00	0.59
35		dom	-7.165e+02	4.3e-01	0.55	-0.26	0.63	1.30
37	SNP479	add	-6.199e+02	9.9e-01	0.33	0.00	1.00	0.59
38		dom	-6.194e+02	4.0e-01	0.32	0.34	0.67	1.28
40	SNP497	add	-7.080e+02	9.8e-01	0.56	0.03	1.00	0.60
41		dom	-7.076e+02	7.1e-01	0.56	0.09	0.64	1.29
43	SNP491	add	-4.041e+02	3.0e-01	0.99	0.62	0.97	0.61
44		dom	-4.041e+02	4.4e-01	0.99	0.13	0.67	1.28
46	SNP459	add	-6.893e+02	9.9e-01	0.50	0.00	1.00	0.59
47		dom	-6.870e+02	**3.7e-02**	0.52	-0.59	0.73	1.25
49	SNP512	add	-6.089e+02	9.9e-01	0.78	0.00	1.00	0.59
50		dom	-6.085e+02	4.7e-01	0.79	-0.31	0.74	1.24

Table 9.4: Data analysis for haplotypes of the Alzheimer's data set. LR is the likelihood ratio statistics and p the p-value based on a χ^2_3 approximation. Parameter estimates are given under the alternative hypothesis. Significant p-values are given in bold.

	SNP	model	p-value	η_1	η_2	η_3	R_1	R_2	R_3	$p^\star$	β
5	SNP886, SNP988	dom	**9.1e-04**	0.375	0.228	0.320	-0.559	-0.254	0.768	0.444	1.462
8	SNP988, SNP888	dom	**6.8e-04**	0.254	0.027	0.314	-0.384	-0.155	-0.378	0.468	1.434
11	SNP888, SNP873	dom	**4.3e-08**	0.103	0.686	0.211	-0.234	0.490	-0.383	0.718	1.264
14	SNP873, SNP952	dom	**1.2e-03**	0.389	0.253	0.358	0.702	-0.358	-0.389	0.420	1.432
17	SNP952, SNP528	dom	**6.8e-05**	0.408	0.007	0.315	0.791	-0.025	-0.592	0.521	1.441
20	SNP528, SNP992	dom	**1.6e-05**	0.008	0.265	0.385	-0.081	-0.297	0.901	0.436	1.392
19	SNP952, SNP992	add	1.0e-00	0.051	0.225	0.669	-0.000	-0.010	0.009	1.000	0.595
20	SNP952, SNP992	dom	5.7e-01	0.053	0.231	0.661	0.008	-0.560	0.501	0.678	1.283
21	SNP952, SNP992	rec	1.0e-00	0.051	0.225	0.669	-0.002	-0.001	0.008	1.000	1.189

9.2.2. *Haplotype analysis.* Results of a two-locus haplotype analysis are reported in table 9.4. The strategy was to slide a window of two SNPs over the available SNPs ordered according to physical position. For locus combinations $(SNP1006, SNP875)$, $(SNP465, SNP457)$, $(SNP471, SNP479)$ and $(SNP479, SNP497)$ Mendelian inconsistencies were discovered for $2, 1, 3, 4$ families, respectively. These errors might be due to genotyping errors, wrong paternity/sample mix-up, but are most likely due to recombinations that formed new haplotypes in offspring. Families with Mendelian inconsistencies were excluded from the analysis. Recombinations are quite likely to be observed in the data set, since markers span over 1 Mb. Implications are discussed below (section 10). Table 9.4 reports only significant findings. The full table is given in the appendix (table D.3, page 187). Again allele frequencies for $p^\star$ and the effect size β agree well. Also there is agreement with the single locus analysis. One exception is model 11, for which $p^\star$ is estimated as 0.72. In contrast to the single locus analysis, significant results span a contiguous region making the localization of the disease gene easier. Apart from the result of model 11 (p $= 4.3e - 8$), SNP $SNP528$ would have been the most likely candidate for the disease gene. Fitting the the model for the haplotype $SNP952$, $SNP992$ which are the loci neighboring $SNP528$ on each side yields unsignificant results for all models, such that omission of $SNP528$ would have made localization more difficult.

In conclusion, results for the dominant model agree well with an independent study [64]. Parameter estimates are consistent with prior knowledge about the ApoE gene, and single and haplotype analysis agree on the allele frequency of the predisposing allele at the disease locus as well as on effect size. Single locus and haplotype analysis complement each other in pinpointing the disease SNP. It would have been interesting to analyze a data set including both SNPs defining the ϵ-alleles (amino acid positions 112 and 158) which are not present in the current dataset. In this case the haplotype analysis should be superior to the single locus analysis.

Table 9.5: Data analysis for a single SNP analysis of the Alzheimer's data set using a random effects model. $\sigma < 10^{-6}$ is reported as 0. l_A is the supremum of the log likelihood under the alternative. The p-value is based on a χ^2_1 approximation. Parameter estimates are given under the alternative.

	SNP	model	l_A	p	η	R	p^*	β	σ
1	SNP1006	add	-673.43	0.99	0.65	4.9e-5	1.00	0.59	0.00
2		dom	-671.73	0.07	0.64	0.82	0.72	1.24	0.00
4	SNP875	add	-667.83	0.99	0.69	1.2e-05	1.00	0.59	0.00
5		dom	-667.26	0.37	0.68	0.72	0.81	1.22	0.00
7	SNP886	add	-658.75	1.00	0.71	0.00	1.00	0.59	0.00
8		dom	-657.97	0.27	0.72	-0.43	0.68	1.27	0.03
10	SNP988	add	-678.13	**0.01**	0.59	-0.62	0.65	0.80	0.03
11		dom	-673.50	**8e-05**	0.59	-0.85	0.49	1.41	0.00
13	SNP888	add	-663.36	**0.02**	0.29	-0.76	0.81	0.69	0.004
14		dom	-662.55	**0.01**	0.28	-0.64	0.57	1.34	0.00
16	SNP873	add	-654.91	**0.02**	0.72	0.76	0.82	0.68	0.00
17		dom	-654.52	**0.01**	0.73	0.59	0.56	1.35	0.00
19	SNP952	add	-669.15	0.72	0.72	0.30	0.96	0.61	0.00
20		dom	-668.36	0.25	0.71	0.49	0.68	1.27	0.00
22	SNP528	add	-683.68	**8e-4**	0.38	0.71	0.54	0.88	0.03
23		dom	-676.57	**7e-7**	0.38	0.89	0.43	1.47	0.007
25	SNP992	add	-646.44	1.00	0.27	-4.59e-05	1.00	0.59	2e-6
26		dom	-644.66	0.07	0.29	-0.88	0.76	1.24	9e-4
28	SNP465	add	-707.15	1.00	0.44	0.00	1.00	0.59	0.00
29		dom	-706.93	0.79	0.44	-0.12	0.72	1.25	4e-5
31	SNP457	add	-720.21	0.83	0.51	-0.18	0.97	0.61	0.00
32		dom	-719.69	0.41	0.52	-0.25	0.64	1.30	0.00
34	SNP471	add	-717.07	1.00	0.54	0.00	1.00	0.59	0.00
35		dom	-716.52	0.43	0.55	-0.26	0.64	1.30	2e-4
37	SNP479	add	-619.87	1.00	0.33	0.00	1.00	0.59	0.00

38		dom	-619.41	0.41	0.32	0.34	0.67	1.28	4e-6
40	SNP497	add	-708.02	0.98	0.56	0.09	0.99	0.60	0.00
41		dom	-707.63	0.72	0.56	0.09	0.64	1.29	0.00
43	SNP491	add	-403.94	0.26	0.99	1.00	0.99	0.60	0.00
44		dom	-404.10	0.44	0.99	0.14	0.69	1.27	0.00
46	SNP459	add	-689.30	1.00	0.50	0.00	1.00	0.59	0.00
47		dom	-686.33	0.02	0.51	-0.72	0.65	1.27	0.03
49	SNP512	add	-608.90	1.00	0.78	3.99e-05	1.00	0.59	0.00
50		dom	-608.37	0.41	0.79	-0.37	0.67	1.28	0.00

10. Discussion

In this thesis, a likelihood framework is developed and characterized that allows to identify important association parameters in genetic association studies. The potential of this model lies in the fact that as yet unexplored parameters are estimated that can shorten the time of gene identification. This framework is developed rigorously for a subset of models, showing asymptotic consistency of estimates. Finally, the likelihood framework is used both in a frequentist and Bayesian setting, exploiting the mutual strengths of the approaches.

10.1. **Robustness.** Robustness in genetic studies concerns mainly two factors: misspecification of the penetrance model and violation of Hardy-Weinberg equilibrium, the latter of which should mainly result from population stratification. Population stratification is not considered in this thesis. Although population stratification was shown to exist in populations [103, 8], there are many reasonable methods to control for stratification [3, 16, 80, 80, 110]. For example, sub-population membership probabilities can be estimated and used to stratify the likelihood. Another issue is misspecification of the mode of inheritance. The methods presented here turn out to be more powerful, yet equally robust as non-parametric methods for testing. Parameter estimates are biased under misspecification. However, often the bias is small and interpretations remain valid. In conclusion, the likelihood framework offers a reasonably robust way to estimate and test parameters which harbor important information needed in the scientific process of clarifying gene functions.

10.2. **Guidance of experimental design.** The methods presented in this thesis allow to identify SNPs that are associated with a binary phenotype. The ongoing experimental design can then be optimized by parameter estimates. For example, an interesting resource has been created by the international HapMap project [13]. The goal of this project was to characterize the joint allele distributions in several important populations. This database can be used to search for SNPs that match the characteristics identified by parameter estimates. Both, the allele frequency of the potential disease allele and LD with observed alleles can be checked against the database. If functions of alleles identified by this process are known, the penetrance parameter can also be used in comparisons and plausibility checks.

10.3. **Ascertainment.** One drawback of the methods in this thesis is ascertainment. The problem lies in the requirement that if ascertainment is on phenotypes of k individuals, $k+2$ individuals have to be sampled per nuclear family. Also the two extra individuals have to be related. This precludes the very common trio design with one affected offspring, since parents are assumed to be unrelated. In this instance at least one more sib has to be ascertained irrespective of phenotype.

However, with the advent of research in normal traits, population based samples should become more common. Without ascertainment on phenotypes, sib pairs represent a convenient and efficient sampling scheme.

10.4. **Haplotype effects.** It might seem that restricting $G^\star$ to a single locus could seriously limit applications, since haplotype effects could be relevant in the disease process, rather than single loci. However $G^\star$ can be considered to be an indicator variable on any causative factor. This factor could be a haplotype. Of course, genotyping only a single SNP would not give enough information about haplotypes in LD with this SNP. If, however, a haplotype analysis is performed, each of the haplotypes has a correlation structure with $G^\star$ allowing to capture haplotypes. Ultimately, the causative haplotype itself would be genotyped and $G^\star$, interpreted as indicator variable, on its own occurrence would show correlation 1.

10.5. **Limitations.** The methods presented here are better suited for tightly linked markers. The data set analyzed, contained widely spaced markers that exhibited recombinations. These are currently not heeded in the model and exclusion of families can lead to biases in parameter estimates and violations of α-levels [88]. Another limitation are requirements on family structures as discussed above (section 10.3). These are more restrictive than those for the TDT/FBAT tests, although in many practical situations ascertainment should be possible.

10.6. **Biological relevance.** It is interesting to look at the biological level to understand the implications of the models suggested in this thesis. In general, it is even hard to give a precise definition of a gene [57]. A common idea is that a segment of DNA is called a gene if in some cells this segment is copied (transcribed) to RNA, which in turn modifies cellular functions. Genes can be modified by regulatory elements not contained in the gene in such a way that polymorphisms cannot easily be judged with respect to their functional relevance. Statistical measures like the correlation used in this thesis therefore are only a first step to characterize gene functions and the relevance of polymorphisms.

10.6.1. *Association versus causation of genetic factors.* The term association is used to denote predictors which have a statistically significant value to predict features in individuals (in the epidemiological context). In genetic studies, such a predictor is given by the penetrance function. However, for a given penetrance function it is unclear, whether it represents the "best" predictor and whether it represents causative effects in the intuitive sense of a biological effect. The idea in this thesis is to use information on related individuals to define causation of alleles/haplotypes by means of a correlation. It is intuitively clear that phenotypes of related individuals are positively correlated, since the occurrence of genotypes is positively correlated in these individuals. On the other hand, if all causative alleles are observed, the phenotype of an individual should only depend on these variations and should be independent in related individuals. In summary, the conditional correlation of phenotypes, given genotypes should be zero if and only if all causative loci for a given phenotype are being observed. However, in the case of perfect LD, *i.e.* any allele on a haplotype allows for a certain prediction of alleles at other loci on that haplotype, the biological effect of these alleles cannot be distinguished. Consider Figure 10.6.1 where three SNPs exist in a region which are all equally informative for two possible haplotypes. In the epidemiological sense a counterfactual argument[11] would prove association of one of the haplotypes as we do not have to remove haplotypes which mix "A" with "a" alleles as they do not exist. However, if say SNP1 is unobserved, it could still have the biological effect by another counterfactual argument: changing SNP1 physically on the DNA could change the phenotype. The problem is, that these haplotypes cannot be observed by assumption, yet it would be crucial, say in drug development, to identify the biologically causative SNP. On the one hand this implies that actual laboratory experiments are needed to determine biological causation, on the other hand different populations could be used as another statistical approach to narrow down on causation since LD varies across populations (by definition). This might lead to the distinction of the following genotype effects:

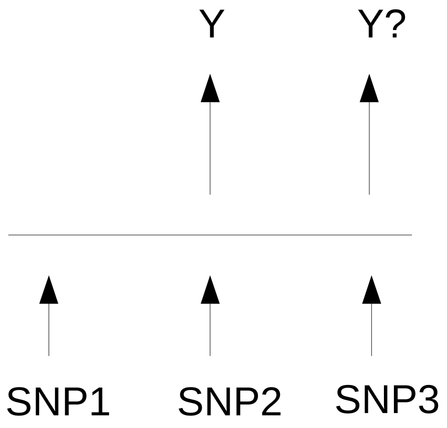
- Association: $E(Y|G) \neq \mu := E(Y)$, *i.e.* the genotype is a predictor of the phenotype
- Statistical causation: $Cov(Y_1, Y_2|G_1, G_2) = 0$ for two related individuals, *i.e.* the observed genotypes solely predict the phenotype
- Biological causation: The experimental change of a causative allele changes the phenotype

Of course, model misspecification and other errors contribute to biases in the parametric model presented here. This makes it highly problematic to determine causation on the basis of parameter estimates and tests. However, it is interesting to see that at least theoretically, by closely modeling the genetic situation, concepts close to causation can be established and can help to actually advance the understanding of biological causation.

10.7. **Genome wide association scans.** The HapMap [13] was an effort to characterize the haplotype distribution in human genomes and has paved the way for a comprehensive analysis of SNPs covering roughly 70% of genetic variation. A practical limitation of this model is that the number of parameters grows exponentially with the number of loci

[11] Setting all possible confounders to a defined state: assume at certain loci a certain allele is observed and allow variation only at loci of interest

Figure 10.1: Figure to help illustrate counterfactual ideas in the genetic context. Three SNPs are observed which only exhibit two haplotypes, *i.e.* there is perfect linkage disequilibrium between the SNPs. The phenotypic effect is in question.



Ht	Alleles	Freq
1	A/A/A	0.5
2	a/a/a	0.5

included in the analysis. In the data analysis presented here I first considered each locus individually, and followed up with haplotypes comprised of two loci. Chen & Abecasis [10] proposed a similar two-stage approach for a model of genotype imputation in families on a whole genome basis. As a general guideline, this model could be used with individual loci to reassess the association found in a prior screening stage. Models with and without random effect could be compared to assess which level of model complexity is needed. The next steps would then include sliding window analyses for two or three SNPs to pinpoint haplotypes that show associations with the disease allele. Finally, it can be worthwhile to analyze non-adjacent SNPs jointly in a haplotype analysis, if there is evidence for a combined effect, such as parameter estimates from previous stages, LD patterns or prior knowledge. To correct for multiple testing, methods based on p-values [35, 32, 36, 33, 5] may be most practical, as the use of permutation methods is difficult due to the computational burden of this model. A sliding window approach suggested for genome wide association studies (GWASs) and case-control data [39] uses a permutation procedure that evaluates a $\min p$-statistic based on the score function which could be computed efficiently in this setting. In principle, this approach is applicable and may warrant further investigation.

10.8. **Future work.** Obviously, it would be interesting to analyze data sets for which the model in this thesis should provide additional information on top of previous analyses. This includes true haplotype effects, covariates and tightly linked markers in data sets. Also the effect of population stratification would be worthwhile to consider. These analyses could help to further evaluate limitations and possibilities of the model in real world applications.

10.8.1. *Model extensions.* The methods of this thesis allow for multiple extensions. It would be interesting to include multiple observed regions into the model, that might each be associated with a disease locus and which might interact with each other. It is possible to integrate recombinations into the model; however, numerical problems are likely to occur. Therefore, some simple and robust method - like associating recombining families with weights - could be developed. Quantitative traits can straightforwardly be implemented and would offer an application of the likelihood for randomly sampled families, if population based samples are analyzed.

10.8.2. *Computational issues.* The computational elegance of the MCMC sampler can be complemented by a re-implementation of the code in C to provide for efficient computations. Also robustness can be increased, esp. in the case of misspecification of the mode of inheritance. In some cases acceptance probabilities of the Metropolis-Hastings/Metropolis samplers were to high to allow for fast convergence. The proposal distributions can be chosen appropriately to achieve rejection rates of 20% - 40%. Another strategy would be to change the current block-updating scheme to a scheme that alternates between updating part of the H^* and β several times during one cycle.

11. Acknowledgments

I would like to thank Ruth Pfeiffer for support and for inviting me to come to the Biostatistics Branch of the National Cancer Institute, USA. On many occasions she offered timely suggestions that allowed me to surmount many problems. I am grateful to Eden R. Martin for providing the data set and to Beate and Claus Blecher for reading the manuscript and checking the language. The high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov) were used for simulations. Greetings and thanks go to Prof. Dr. Lohmann for his critical appraisal of the manuscript.

Appendix A. Abbreviations and Glossary

Table A.1: Abbreviations

A	the nucleotide Adeonsine
AC	Allele Combination
ApoE	Apolipoprotein E
ASP	Affected Sib Pair
bp	base pair(s)
C	the nucleotide Cytidine
CDF	Cumulative Distribution Function
DNA	Deoxyribo Nucleic Acid
EM algorithm	Expectation Maximization algorithm
fbat	family based association test
G	the nucleotide Guanosine
GWAS	Genome wide association study
HWE	Hardy-Weinberg Equilibrium
IBD	Identical By Descent
IBS	Identical By State
Kb	kilo base pairs
LD	Linkage disequilibrium
Mb	mega base pairs
MCMC	Marcov Chain Monte Carlo
MIM	Mendelian inheritance in men (code of identification)
ML	Maximum likelihood
MLE	Maximum likelihood estimator
MS	Multiple Sclerosis
MSE	Mean Squared Error
OR	Odds ratio
PCR	Polymerase chain reaction
RNA	Ribo Nucleic Acid
RV	Random Variable
SNP	Single Nucleotide Polymorphism
T	the nucleotide Thymidine
TDT	Transmission Disequilibrium Test
VNTR	Variable Number of Tandem Repeats
WGA	Whole genome analysis
WLOG	without loss of generality

Table A.2: Glossary

Allele	Sequence occurring at a locus
Allele dose	the number of one designated allele in a genotype which can assume values $0, 1, 2$
base	chemical component that constitutes the DNA (apart from linking or backbone components)
base pair	two bases that are linked by weak chemical bonds and are part of different DNA strands. Only two different base pairs occur, namely A/T, G/C
bi-allelic Locus	Polymorphism for which two alleles can be observed
Genetic map	Distance between two markers defined by the recombination fraction between markers
Locus	Physical location in the genome defined by start and extent
Polymorphism	Marker for which at least two alleles have frequencies $\geq 1\%$
Marker	Locus for which at least two alleles can be observed in the population
Mode of inheritance	Interaction pattern of alleles at a locus that influence a phenotype
Mutation	Event that changes the DNA sequence in a cell
Phase	The state of two alleles at two different loci, being either on the same strand (cis) or on different strands (trans)
Kinship coefficient	Probability that a given locus is shared IBD between two individuals
Physical map	Positions of markers given in base pairs (compare genetic map)

Appendix B. Haplotype reconstruction

The likelihood of an unphased data can be computed by defining a function from the space of diplotypes to the set of genotypes which destroys phase information. This mapping π_G can be defined as a projection, essentially ordering alleles at each locus independently thereby destroying phase information. In section 4 sets of diplotypes and genotypes are defined to have different structure which implies that formally π_G is not a projection but the intuition of a projection helps to guide through the algorithm.

B.1. C code for construction of $\mathcal{H}(G)$ in nuclear families.

```
/*
  geno2haplo.c
  Tue 28 Mar 2006 11:15:06 AM EST
 */


#include  "geno2haplo.h"
#include  <stdlib.h>
#include  <string.h>
#include  <assert.h>
#include  "bitManip.h"
#include  <stdio.h>

/*
  <!> bit operations are sensitive to haptotype encoding (word-size)
 */

/*
  debugging fuctions
 */

void  printGenotypesPlain(int *gts, int countOfLoci, int countOfIndividuals) {
  int  i, j;
  printf("Genotypes: ");
  for (i = 0; i < countOfIndividuals; i++) {
    printf("(");
    for (j = 0; j < countOfLoci; j++) {
      printf("%s%d, %d", j? ", ": "",
        gts[i * (countOfLoci * 2) + 2 * j ], gts[i * (countOfLoci * 2) + 2 * j + 1]);
    }
    printf(") ");
  }
  printf("\n");
}

void  printGenotypes(array_t *gs, int countLoci) {
  int      i, count /* offspring count */ = arrayCount(gs);

  printf("Gts tb reconstr:");
  for (i = 0; i < count; i++) {
    printf(" (%d %d)", GENOTYPE(gs, i)->gts[0], GENOTYPE(gs, i)->gts[1]);
  }
  printf("\n");
}

#define  PRINT_PDTS_OGTS  \
printf("parental:%d-%d %d-%d gts:%d-%d %d-%d %s\n", cmd.hts[0], cmd.hts[1], cpd.hts[0],
  GENOTYPE(gs, 2)->gts[0], GENOTYPE(gs, 2)->gts[1],\
  GENOTYPE(gs, 3)->gts[0], GENOTYPE(gs, 3)->gts[1],\
  (k < oc)? "incompatible": "compatible"\
);


/*
  conversion functions

  <A> missing data convention
  we assume genotypes to be present/absent completely
  a -1 for the first allele inidicates missingness
 */

array_t  *genotypesFromArray(int *gts, int countOfLoci, int countOfIndividuals) {
  array_t  *cgts = arrayWithSizeCount(sizeof(genotype_t), countOfIndividuals, countOfInd
  assert(countOfLoci <= 32);
  int  i, j;

  for (i = 0; i < countOfIndividuals; i++) {
    diplotype_t  d = (diplotype_t) { 0, 0 };
    if (gts[i * (countOfLoci * 2)] == -1) {
      *(genotype_t *)arrayElementAt(cgts, i) = (genotype_t) { -1, -1 };
    } else {
      for (j = 0; j < countOfLoci; j++) {
        if (gts[i * (countOfLoci * 2) + 2 * j ])  SetBitAtL(&d.hts[0], j);
        if (gts[i * (countOfLoci * 2) + 2 * j + 1])  SetBitAtL(&d.hts[1], j);
      }
      *(genotype_t *)arrayElementAt(cgts, i) = genotypeFromDiplotype(d);
    }
  }
  return cgts;
}


/*
  construct diplotypes, compatible with a genotype
  Algorithm:
    for the 2nd to last heterocygous site produce all combinations of phases

  Here are the fine parts of the algorithm:
    - we produce unordered haplotypes (i.e., heterozygotes produce only one instance)
    - in case of missing loci and homocygous genotype the first missing locus
      only contributes 3 possibilities, subsequent missing loci contribute all 4 possibi
    - we use coding trick to avoid special cases (A)
      we code missing genotypes as ( a1, ~ a2 ), such that two-bit codes
      0, 1, 2, 3 code for ordered genotypes 01, 00, 11, 10, respectively
      in case of homocygous genotypes and missing loci the last missing locus only contr
      combinations 0, 1, 2
 */

static void  printReconstruction(array_t *fdts) {
  int i;
  for (i = 0; i < arrayCount(fdts); i++) {
    diplotype_t  *ds = arrayElementAt(fdts, i);
    printf("Maternal haplotype: (%2d, %2d), paternal haplotype: (%2d, %2d)\n",
      ds[0].hts[0], ds[0].hts[1], ds[1].hts[0], ds[1].hts[1]);
    int j;
    for (j = 0; j < offspringCount(fdts); j++) {
      printf("\tOffspring(%d): (%2d, %2d), (%2d, %2d), (%2d, %2d), (%2d, %2d)\n", j,
        ds[2 + j * 4 + 0].hts[0], ds[2 + j * 4 + 0].hts[1],
        ds[2 + j * 4 + 1].hts[0], ds[2 + j * 4 + 1].hts[1],
        ds[2 + j * 4 + 2].hts[0], ds[2 + j * 4 + 2].hts[1],
        ds[2 + j * 4 + 3].hts[0], ds[2 + j * 4 + 3].hts[1]
      );
    }
  }
}

array_t  *geno2haploMissing(genotype_t g, int locusCount, int countMissing) {
  array_t    *dts;
  int      i, j;

  // missing data is expanded to unordered diplotypes
  if (g.gts[0] == -1) {
    int  countHts = 1 << (locusCount + countMissing);
    dts = arrayWithSizeCount(sizeof(diplotype_t), 0, countHts * (countHts - 1));
    for (i = 0; i < countHts; i++) {
      for (j = i; j < countHts; j++) {
        diplotype_t  *d = arrayAddSlot(dts);
        *d = (diplotype_t){ i, j };
      }
    }
    return dts;
  }

  HetCount_t  hc = countHets(g, locusCount);
  // count of haplotypes due to heterozygosity
  int      hetBits = (hc.count <= 1? 0: (hc.count - 1));
  // this is an upper limit for the amount of haplotypes (observed hom, unobserved het).
  int      countDts = (1 << hetBits) * 1 << (2 * countMissing);

  if (hc.count == 0 && countMissing > 0) {    /* case A from above */
    countDts >>= 2, countDts *= 3;
  }
  dts = arrayWithSizeCount(sizeof(diplotype_t), 0, countDts);

  for (i = 0; i < countDts; i++) {
    diplotype_t  *d = arrayAddSlot(dts);
    *d = (diplotype_t){ g.gts[0], g.gts[1] };  // pre-initialize homocygous positions
                          // (and last heterozygous position)
//printf("gt:(%d, %d)", d->hts[0], d->hts[1]);
    for (j = 0; j < hetBits; j++) {
      SetBitAtToL(&d->hts[0], hc.positions[j],  BitAtL(&i, j));
//printf(" bit:%d ", BitAtL(&i, j));
      SetBitAtToL(&d->hts[1], hc.positions[j], !BitAtL(&i, j));
    }
//printf("gt:(%d, %d)", d->hts[0], d->hts[1]);
    for (j = 0; j < countMissing; j++) {
//printf("i:%d @%d: Bit1:%d Bit2:%d (%d, %d)", i, j, BitAtL(&i, hetBits + 2 * j), !BitAt
      SetBitAtToL(&d->hts[0], locusCount + j,  BitAtL(&i, hetBits + 2 * j));
      SetBitAtToL(&d->hts[1], locusCount + j, !BitAtL(&i, hetBits + 2 * j + 1));
//printf(" (%d, %d)\n", d->hts[0], d->hts[1]);
    }
  }
  return dts;
}

array_t  *geno2haplo(genotype_t g, int countLoci) {
  return geno2haploMissing(g, countLoci, 0);
}

/*
  build unordered (ie sorted) representation of genotypes, ie we have to sort two word b
 */
```

```
genotype_t  genotypeFromDiplotype(diplotype_t d) {
  // heterocygous loci, ie these positions have to be sorted
  haplotype_t  hets = d.hts[0] ^ d.hts[1];
  // ordering is naturally given by bit operations, setting the ’first’ genotype to 0 at
  // and to 1 for the ’second’ genotype
  return (genotype_t) { 0 | (~hets & d.hts[0]), hets | (~hets & d.hts[0]) };
}

/*
  reconstruct haplotpes in a nuclear family
  - iterate combinations of parental diplotypes
    - iterate offspring to check for compatibility (optimization: order by increasing am
    - there are at most 2 possible transmission patterns for parental diplotypes:
      - assume an offspring diplotype, compatible
      - by subtraction of one compatible haplotype (parent-offspring) the other halptoyp
        otherwise for the first diplotype had to differ from its current state (to allow
      - corollary: if we have found one compatible offspring diplotype the search can be

  Data Structure: parental diplotypes, two diplotypes per offspring (one is voided in ca
 */

static inline int  cmpGenotypes(genotype_t g0, genotype_t g1, int countLoci) {
  int  mask = (1 << countLoci) - 1;
  int cmp1 = (g1.gts[0] & mask) - (g0.gts[0] & mask);
  return cmp1? cmp1: ((g1.gts[1] & mask) - (g0.gts[1] & mask));
}

array_t  *geno2haploFamMissing(array_t *gs, int countLoci, int countMissing) {
  // maternal, paternal diplotypes
  array_t  *md = geno2haploMissing(*GENOTYPE(gs, 0), countLoci, countMissing),
      *pd = geno2haploMissing(*GENOTYPE(gs, 1), countLoci, countMissing);
  int      i, j, k, mx, oc /* offspring count */ = arrayCount(gs) - 2;
  BOOL    doRetain;
  array_t    *r = arrayWithSizeCount(sizeof(diplotype_t) * (2 + 4 * oc), 0,
    arrayCount(md) * arrayCount(pd));  // reconstruction
  diplotype_t  *cr;  // current return

  //printGenotypes(gs, countLoci);
  // reconstruct diplotypes per individual

  // iterate mother
  for (i = 0; i < arrayCount(md); i++) {
    diplotype_t  cmd = *DIPLOTYPE(md, i);
    // iterate father
    for (j = 0; j < arrayCount(pd); j++) {
      diplotype_t  cpd = *DIPLOTYPE(pd, j);
      cr = arrayAddSlot(r);
      assert(cr != 0);
      cr[0] = cmd, cr[1] = cpd;
      // iterate offspring compare genotypes for parental transmissions with offspring g
      //printf("Mat:%d, %d, Pat: %d, %d\n", cmd.hts[0], cmd.hts[1], cpd.hts[0], cpd.hts[
      for (k = 0; k < oc; k++) {
        diplotype_t  cod;
        BOOL      c00, c01, c10, c11;
        genotype_t  og = *GENOTYPE(gs, 2 + k);
        cr[2 + 4 * k] = cr[2 + 4 * k + 1] = cr[2 + 4 * k + 2] = cr[2 + 4 * k + 3] = NA_D
        // we allow for duplicates to simplify the algorithm
        // we place restrictions only on observed genotypes
        cod = (diplotype_t){cmd.hts[0], cpd.hts[0]};
        if (c00 = !cmpGenotypes(og, genotypeFromDiplotype(cod), countLoci))
          cr[2 + 4 * k] = cod;

        cod = (diplotype_t){cmd.hts[0], cpd.hts[1]};
        if (c01 = !cmpGenotypes(og, genotypeFromDiplotype(cod), countLoci))
          cr[2 + 4 * k + 1] = cod;

        cod = (diplotype_t){cmd.hts[1], cpd.hts[0]};
        if (c10 = !cmpGenotypes(og, genotypeFromDiplotype(cod), countLoci))
          cr[2 + 4 * k + 2] = cod;

        cod = (diplotype_t){cmd.hts[1], cpd.hts[1]};
        if (c11 = !cmpGenotypes(og, genotypeFromDiplotype(cod), countLoci))
          cr[2 + 4 * k + 3] = cod;

        if (!c00 && !c01 && !c10 && !c11) break;  // no compatible diplotype found
      }
      //PRINT_PDTS_OGTS
      if (k < oc) { // no compatible diplotype found
        arrayFreeSlot(r);
      }
    }
  }
  freeArray(md);
  freeArray(pd);
  //printReconstruction(r);
  return r;
}

array_t  *geno2haploFam(array_t *gs, int countLoci, int countMissing) {
  return geno2haploFamMissing(gs, countLoci, 0);
}
```

Appendix C. Appendix simulations

C.1. **Implementation of the likelihood.** The likelihood is implemented in C for efficient execution. An R interface allows to use the function with R. Analogous functions exist for the random effects likelihood.

```
/*
    likelihood-1.c
    Fri 21 Apr 2006 11:10:23 AM EDT
 */

#include    "geno2haplo.h"
#include    <math.h>
#include    "bitManip.h"
#include    <stdio.h>
#include    "parameterization.h"

/* <par> log-likelihood for single region */

/*
    likelihood for a single family
 */

/*
    we expect data structures mangled through R
 */

#define    MAX(a, b)    ((a) < (b)? (b): (a))

/*
    debugging macros
 */
#if !defined(_NDEBUG)
#    define    PRINT_PARS_R2(cLd, beta, countLoci)    printParsR2(cLd, beta, countLoci)
#    define    PRINT_PARS_HF(hfs, beta, countLoci) printParsHf(hfs, beta, countLoci)
#    define    PRINT_LL(name, ll)    printf("Log likelihood (%s): %.4e\n", name, ll)
#elif
#    define    PRINT_PARS_R2(cLd, beta, countLoci)    (0)
#    define    PRINT_PARS_HF(hfs, countLoci)    (0)
#    define    PRINT_LL(ll)    (0)
#endif

// <A> disease locus is adjecent to observed loci (== countLoci)
static inline double    penetranceFunctionDominant(double beta, int y, int *dt, int coun
    double baseline) {
    int    g[2] = { BitAt(dt + 0, countLoci), BitAt(dt + 1, countLoci) };
    int    score = MAX(g[0], g[1]);    // dominant model
    double    logitEffect = baseline + beta * score;
    double    logisticProb = exp(logitEffect) / (1 + exp(logitEffect));
    return y? logisticProb: (1 - logisticProb);
}
static inline double    penetranceFunctionRecessive(double beta, int y, int *dt, int cou
    double baseline) {
    int    g[2] = { BitAt(dt + 0, countLoci), BitAt(dt + 1, countLoci) };
    int    score = g[0] && g[1];      // Recessive model
    double    logitEffect = baseline + beta * score;
    double    logisticProb = exp(logitEffect) / (1 + exp(logitEffect));
    return y? logisticProb: (1 - logisticProb);
}
static inline double    penetranceFunctionAdditive(double beta, int y, int *dt, int coun
    double baseline) {
    int    g[2] = { BitAt(dt + 0, countLoci), BitAt(dt + 1, countLoci) };
    int    score = g[0] + g[1];    // additive model
    double    logitEffect = baseline + beta * score;
    double    logisticProb = exp(logitEffect) / (1 + exp(logitEffect));
    return y? logisticProb: (1 - logisticProb);
}

/*
    each haplotype configuration is of identical structure:
    parents, 4 configurations per offspring

 */
#define    DT_P(config, parent)    (fdts + ((config) * (offspringCount * 2 * 4  + 4)) +
#define    DT_O(config, offspring)    (fdts + ((config) * (offspringCount * 2 * 4  + 4))
#define    LH_DT(dt)    (hf[dt[0]] * hf[dt[1]] * (dt[0] == dt[1]? 1: 2))

#define    DEFINE_L_1_F(name, PenetranceFunction) \
static inline double    l_1_f_##name(double *pars,int *fdts, int countConfigs, int *phen
    int offspringCount,    int countLoci, double baseline) { \
    int        i; \
    int        counthf = 1 << (countLoci + 1); \
    double    fl = 0;    /* family likelihood */ \
    double    beta = pars[counthf]; \
    double    *hf = pars; \
    \
    /*PRINT_PARS_HF(pars, beta, countLoci); */\
    for (i = 0; i < countConfigs; i++) {\
        double    l;    /* family likelihood */\
        int        *mdt = DT_P(i, 0), *pdt = DT_P(i, 1); \
        l = LH_DT(mdt) * LH_DT(pdt) \
            * PenetranceFunction(beta, phenotypes[0], mdt, countLoci, baseline) \
            * PenetranceFunction(beta, phenotypes[1], pdt, countLoci, baseline);\
        double    mf = 0.25; \
        int j; \
        double    ol;    /* offspring likelihood */ \
        for (j = 0; j < offspringCount; j++) { \
            int    *odt = DT_O(i, j); \
            int k; \
            ol = 0; \
            /* sum over all four possible transmissions as indicated in the data structu
            for (k = 0; k < 4; k++) { \
                /* printf("offspring:%.3e %d %d %d\n", ol, odt[2*k], odt[2*k+1], phenoty
                if (odt[2 * k] >= 0) { \
                    ol += PenetranceFunction(beta, phenotypes[2 + j], odt + 2*k, countLo
                } \
            } \
            l *=  mf * ol; \
        } \
        fl += l; \
    } \
    /*PRINT_LL("family", fl);*/ \
    return fl; \
}

/*
    for now we only allow for missing parental phenotypes
    we marginalize over the missing phenotype
    missing genotypes are handled through the reconstruction
 */

#define    DEFINE_LL_1_F_MP(name, PenetranceFunction) \
DEFINE_L_1_F(name, PenetranceFunction) \
static inline double    ll_1_f_mp_##name(double *pars,int *fdts, int countConfigs, int *
    int offspringCount,    int countLoci, double baseline) { \
    double    l = 0; \
    int    pm1, pm2, pp1, pp2, i; \
    int    impPhenotypes[offspringCount + 2];    /* imputed phenotypes */ \
    \
    for (i = 0; i < offspringCount; i++) impPhenotypes[i + 2] = phenotypes[i + 2]; \
    \
    if (phenotypes[0] < 0) pm1 = 0, pm2 = 1; \
    else pm1 = pm2 = phenotypes[0]; \
    \
    if (phenotypes[1] < 0) pp1 = 0, pp2 = 1; \
    else pp1 = pp2 = phenotypes[1]; \
    \
    for (; pm1 <= pm2; pm1++) { \
        for (; pp1 <= pp2; pp1++) {    \
            impPhenotypes[0] = pm1; \
            impPhenotypes[1] = pp1; \
            l += l_1_f_##name(pars, fdts, countConfigs, impPhenotypes, offspringCount, c
        } \
    } \
    return log(l); \
}

// Mendelian probability that allele 0 is transmitted from gt
#define    MENDEL0(gt)    (!(gt)? 1: ((gt) == 1? 0.5: 0))
#define    MENDEL(a, gt)    ((a)? (1 - MENDEL0(gt)): MENDEL0(gt))

// conditioning likelihood component on one affected child
// parameterized by r2/ld (completed)

#define    DEFINE_LL_ASC_OAO_R2(name, PenetranceFunction)\
static double    ll_asc_r2_##name##_oao(double *pars, int countLoci, int countOffspring,
    static int    dts[3][2] = {{0, 0}, {0 , 1}, {1, 1}};\
    int        counthf = 1 << (countLoci + 1);\
    double    ps = pars[counthf];    /* disease allele frequency */\
    double    beta = pars[counthf + 1]; \
    double    l = 0, ol; \
    int    pg, mg; \
    double    pGts[3] = { (1 - ps) * (1 - ps), 2 * (1 - ps) * ps, ps * ps }; \
    \
    for (mg = 0; mg <= 2; mg++) { \
        for (pg = 0; pg <= 2; pg++) { \
            ol = 0; \
            /* <A> our artificial diplotypes have the disease locus at position 0 */\
            if (mg < 2 && pg < 2) \
                ol += PenetranceFunction(beta, 0, dts[0], 0, baseline) * MENDEL(0, mg) *
            \
            if (!((mg == 2 && pg == 2) || (mg == 0 && pg == 0))) \
                ol += PenetranceFunction(beta, 0, dts[1], 0, baseline) \
                    * (MENDEL(0, mg) * MENDEL(1, pg) + MENDEL(1, mg) * MENDEL(0, pg)); \
            \
            if (mg > 0 && pg > 0) \
                ol += PenetranceFunction(beta, 0, dts[2], 0, baseline) * MENDEL(1, mg) *
            \
            /* given paternal genotypes, offspring genotypes are mendelian */\
            l += pow(ol, countOffspring) * pGts[mg] * pGts[pg]; \
        } \
    } \
    return log(1 - l); \
}

/*
    pars: \beta, \eta: conditional folding
 */
```

```
#define    DEFINE_LL_1(name, PenetranceFunction) \
DEFINE_LL_1_F_MP(name, PenetranceFunction) \
double    ll_1_##name(double    *pars, int *fdts, int *phenotypes, \
    int countFams, int *countConfigs, int *familySizes,  int countLoci, double baseline)
    double    ll = 0; \
    int i, foffset, poffset; \
    \
    for (i = foffset = poffset = 0; i < countFams; i++) { \
        ll += ll_1_f_mp_##name(pars, \
            fdts + foffset, countConfigs[i], \
            phenotypes + poffset, familySizes[i] - 2, \
            countLoci, baseline \
        ); \
        foffset += countConfigs[i] * (4 + 8 * (familySizes[i] - 2));  \
        poffset += familySizes[i]; \
    } \
    return ll; \
}

#define    SMALL_LL    (-1e8)

#define    DEFINE_LL_1_C(name, PenetranceFunction) \
double    ll_1_c_##name(double *pars, int *fdts, int *phenotypes, \
    int countFams, int *countConfigs, int *familySizes,  int countLoci, double baseline)
    int        counthf = 1 << (countLoci + 1), i; \
    double    newpars[counthf + 1]; \
    double    hf[counthf], hfn[counthf]; \
    double    sumhf; \
    \
    for (i = 0, sumhf = 0; i < counthf - 1; i++) sumhf += newpars[i] = hf[i] = pars[i];
    newpars[counthf - 1] = hf[counthf - 1] = 1 - sumhf; \
    newpars[counthf] = pars[counthf - 1]; \
    \
    if (sumhf > 1) return SMALL_LL; \
    return ll_1_##name(newpars, fdts, phenotypes, countFams, countConfigs, familySizes,
}

// simple constraint management, r2 parameterization
// hfreq: || = 2 ** (cl - 1) - 1, beta: || = 1, diseasep: || = 1, r: || = 2 ** (cl - 1)

static void    printParsR2(double *pars, double beta, int countLoci) {
    int        counthf = 1 << (countLoci + 1), i;
    printf("Pars (LD): hfs = (");
    for (i = 0; i < counthf / 2; i++) {
        printf("%s%.3f", i? ", ": "", pars[i]);
    }
    printf("), r2 = (");
    for (i = counthf/2; i < counthf; i++) {
        printf("%s%.3f", (i > counthf/2)? ", ": "", pars[i]);
    }
    printf("), p = %.3f, beta = %2f\n", pars[counthf], beta);
}
static void    printParsHf(double *pars, double beta, int countLoci) {
    int        counthf = 1 << (countLoci + 1), i;
    printf("Pars (HF): hfs = (");
    for (i = 0; i < counthf; i++) {
        printf("%s%.3f", i? ", ": "", pars[i]);
    }
    printf("), beta = %2f\n", beta);
}

#define HF_EPS    1e-6

#define    DEFINE_LL_1_R2_C(name, PenetranceFunction) \
DEFINE_LL_1(name, PenetranceFunction) \
double    ll_1_r2_c_##name(double *pars, int *fdts, int *phenotypes, \
    int countFams, int *countConfigs, int *familySizes, int countLoci, double baseline)
    int        counthf = 1 << (countLoci + 1), i; \
    double    cLd[counthf]; \
    double    newpars[counthf + 1]; \
    double    ll; \
    double    sumhf; \
    \
    if (!completeLd(pars, counthf, cLd)) return SMALL_LL; \
    ld2hFreq(cLd, counthf, newpars); \
    /* PRINT_PARS_R2(cLd, pars[counthf - 1], countLoci); */\
    /* PRINT_PARS_HF(newpars, pars[counthf - 1], countLoci); */\
    for (i = 0, sumhf = 0; i < counthf; i++) { \
        if (newpars[i] <= HF_EPS || newpars[i] >= 1 - HF_EPS) return SMALL_LL; \
        sumhf += newpars[i];    \
    } \
    if (sumhf > 1) return SMALL_LL;    \
    newpars[counthf] = pars[counthf - 1]; \
    ll = ll_1_##name(newpars, fdts, phenotypes, countFams, countConfigs, familySizes, co
    /* PRINT_LL("total", ll); */ \
    return ll; \
}

/* ascertained log likelihood */
#define    DEFINE_LL_ASC_1_R2_C(name, PenetranceFunction, AscCorr) \
DEFINE_LL_1_R2_C(name, PenetranceFunction) \
double    ll_asc_1_r2_c_##name##_##AscCorr(double *pars, int *fdts, int *phenotypes, \
    int countFams, int *countConfigs, int *familySizes,  int countLoci, double baseline)
    int        counthf = 1 << (countLoci + 1), i; \
    double    cLd[counthf + 2]; \
    double    newpars[counthf + 1]; \
    double    asc = 0; \
    double    sumhf; \
    \
    if (!completeLd(pars, counthf, cLd)) return SMALL_LL; \
    ld2hFreq(cLd, counthf, newpars); \
    for (i = 0, sumhf = 0; i < counthf; i++) { \
        if (newpars[i] <= HF_EPS || newpars[i] >= 1 - HF_EPS) return SMALL_LL; \
        sumhf += newpars[i];    \
    } \
    newpars[counthf] = cLd[counthf + 1] = pars[counthf - 1]; \
    \
    for (i = 0; i < countFams; i++)    \
        asc += ll_asc_r2##_##name##_##AscCorr( \
            cLd, countLoci, familySizes[i] - 2, baseline); \
    \
    return ll_1_##name(newpars, fdts, phenotypes, countFams, countConfigs, \
        familySizes, countLoci, baseline) - asc; \
}

DEFINE_LL_ASC_OAO_R2(dom, penetranceFunctionDominant)
DEFINE_LL_ASC_OAO_R2(add, penetranceFunctionAdditive)
DEFINE_LL_ASC_OAO_R2(rec, penetranceFunctionRecessive)

DEFINE_LL_ASC_1_R2_C(dom, penetranceFunctionDominant, oao)
DEFINE_LL_ASC_1_R2_C(add, penetranceFunctionAdditive, oao)
DEFINE_LL_ASC_1_R2_C(rec, penetranceFunctionRecessive, oao)

DEFINE_LL_1_C(dom, penetranceFunctionDominant)
DEFINE_LL_1_C(add, penetranceFunctionAdditive)
DEFINE_LL_1_C(rec, penetranceFunctionRecessive)
```

C.2. **Estimation of Fisher-Information.** Fisher information is estimated numerically by computing first and second derivatives of the likelihood. The R code for the derivation is adopted from the Numerical Recipes in C [78].

```
# R code for computing the Fisher information
#
#  fisher.R
#Wed 12 Jul 2006 07:11:34 PM EDT

gradient = function(f, pt, h = NULL, eps = 1e-5, ...) {
  vec.dim = length(pt);
  if (is.null(h)) { h = rep(eps, vec.dim) };
  if (is.matrix(h)) h = diag(h);
  sapply(1:vec.dim, function(d){
    delta = rep(0, vec.dim);
    delta[d] = h[d];
    grad = (f(pt + delta, ...) - f(pt - delta, ...)) / (2 * h[d]);
    #print(list(grad =grad, pt = pt, delta = delta, h = h));
    grad
  })
}

# assume h to be a matrix <!> otherwise create one...
hessian.fancy = function(f, pt, h = NULL, eps = 1e-6, ...) {
  vec.dim = length(pt);
  if (is.null(h)) { h = rep(eps, vec.dim) };
  hess = matrix(rep(0, vec.dim **2), vec.dim, vec.dim);
  diag(hess) = sapply(1:vec.dim, function(d){
    delta = rep(0, vec.dim);
    delta[d] = h[d, d];
    ((f(pt + 2 * delta, ...) - f(pt, ...)) -
     (f(pt, ...) - f(pt - 2 * delta, ...))) / (4 * h[d, d]^2)
  })

  for (i in 1:(vec.dim - 1)) {
    for (j in (i + 1):vec.dim) {
      delta1 = rep(0, vec.dim);
      delta1[i] = h[i, j];
      delta2 = rep(0, vec.dim);
      delta2[j] = h[j, i];
      hess[i, j] = hess[j, i] =
        ((f(pt + delta1 + delta2, ...) - f(pt + delta1 - delta2, ...)) -
         (f(pt - delta1 + delta2, ...) - f(pt - delta1 - delta2, ...))) / (4* h[i, j] *
    }
  }
  hess
}

hessian = function(f, pt, h = NULL, eps = 1e-6, ...) {
  vec.dim = length(pt);
  if (is.null(h)) { h = rep(eps, vec.dim) };
  hess = matrix(rep(0, vec.dim **2), vec.dim, vec.dim);
  diag(hess) = sapply(1:vec.dim, function(d){
    delta = rep(0, vec.dim);
    delta[d] = h[d];
    ((f(pt + 2 * delta, ...) - f(pt, ...)) -
     (f(pt, ...) - f(pt - 2 * delta, ...))) / (4 * h[d]^2)
  })
  for (i in 1:(vec.dim - 1)) {
    for (j in (i + 1):vec.dim) {
      delta1 = rep(0, vec.dim);
      delta1[i] = h[i];
      delta2 = rep(0, vec.dim);
      delta2[j] = h[j];
      hess[i, j] = hess[j, i] =
        ((f(pt + delta1 + delta2, ...) - f(pt + delta1 - delta2, ...)) -
         (f(pt - delta1 + delta2, ...) - f(pt - delta1 - delta2, ...))) / (4* h[i] * h[j
    }
  }
  hess
}

invert = function(m) {
  svd = svd(m);
  svd$v %*% diag(1/svd$d) %*% t(svd$u)
}

# <i> define parameters
fisher.family.wise.1 = function(d, mle, likelihood, p, baseline, h = NULL, fisher.eps =
  if (is.null(h)) { h = rep(fisher.eps, length(p$dpars)); }

  counter.dts = 1;
  counter.pts = 1;
  counter.gts = 1;
  cl = d$sim$countLoci;
  dim.par = 2 ** (cl + 1);
  fisher = matrix(rep(0, dim.par * dim.par), c(dim.par, dim.par));
  for (i in 1:length(d$sim$familySizes)) {
    size = d$sim$familySizes[i];
    f.phenotypes = d$sim$phenotypes[counter.pts:(counter.pts + size - 1)];
    count.gts = (cl * 2) * size;
    f.genotypes = d$sim$genotypes[counter.gts:(counter.gts + count.gts - 1)];
    count.dts = 2 * (2 + (size - 2) * 4) * d$reconstruction$configCounts[i];
    f.dts = d$reconstruction$diplotypes[counter.dts:(counter.dts + count.dts - 1)];

#print(list(mle = mle, lh = likelihood, fisher.eps, f.dts, pts = f.phenotypes, d$reconst
    score = gradient(likelihood, mle, h, fisher.eps,
      f.dts, f.phenotypes, d$reconstruction$configCounts[i], c(size), cl, baseline);
#print(score);
    if (!any(is.numeric(score))) {
      print(list(scoreFailed = score));
      return(matrix(1e10, nrow = dim.par, ncol = dim.par));
    }
    fisher.i = score %*% t(score);
    fisher = fisher + fisher.i;
    counter.pts = counter.pts + size;
    counter.gts = counter.gts + count.gts;
    counter.dts = counter.dts + count.dts;
  }
  fisher = fisher / length(d$sim$familySizes);
  fisher;
}

fisher.family.wise.2 = function(d, mle, likelihood, p, baseline, h, fisher.eps = 1e-6) {
  if (is.null(h)) { h = rep(fisher.eps, length(p$dpars)); }
  clo = d$sim$countLoci;

  counter.dts = 1;
  counter.pts = 1;
  counter.gts = 1;
  cl = d$sim$countLoci;
  dim.par = 2 ** (cl + 1);
  fisher = matrix(rep(0, dim.par * dim.par), c(dim.par, dim.par));
  for (i in 1:length(d$sim$familySizes)) {
    size = d$sim$familySizes[i];
    f.phenotypes = d$sim$phenotypes[counter.pts:(counter.pts + size - 1)];
    count.gts = (cl * 2) * size;
    f.genotypes = d$sim$genotypes[counter.gts:(counter.gts + count.gts - 1)];
    count.dts = 2 * (2 + (size - 2) * 4) * d$reconstruction$configCounts[i];
    f.dts = d$reconstruction$diplotypes[counter.dts:(counter.dts + count.dts - 1)];

    factors = c(1, 1e2, 1e1, 1e-1, 1e-2, 1e-3);
    for (f in factors) {
      fisher.2 = -hessian(likelihood, mle, h * f, fisher.eps,
        f.dts, f.phenotypes, d$reconstruction$configCounts[i], c(size), clo, baseline
      );
      if (all(fisher.2 < 1e5)) break;
      print(sprintf("failed for factor %.3e", f));
    }
    fisher = fisher + fisher.2;
#if (any(fisher.2 > 1e5)) print(list("bad fisher" = fisher.2))
#else print(list("good fisher" = fisher.2));

    counter.pts = counter.pts + size;
    counter.gts = counter.gts + count.gts;
    counter.dts = counter.dts + count.dts;
  }
  fisher = fisher / length(d$sim$familySizes);
  fisher
}
```

```
fisher.2 = function(d, mle, likelihood, p, baseline, h, fisher.eps = 1e-6) {
  if (is.null(h)) { h = rep(fisher.eps, length(p$dpars)); }
  clo = d$sim$countLoci;

  if (0) {
  factors = c(1, 1e2, 5e1, 1e1, 5e-1, 1e-1, 5e-2, 1e-2, 5e-3, 1e-3, 1e-4, 1e-5);
  fisher.l = list();
  fisher.max.abs.l = list();
  for (f in factors) {
    s.f = f;
    fisher.2 = -hessian(likelihood, mle, h * f, NULL,
      d$reconstruction$diplotypes, d$sim$phenotypes, d$reconstruction$configCounts,
      d$sim$familySizes, clo, baseline
    );
    fisher.l[[length(fisher.l) + 1]] = fisher.2;
    fisher.max.abs.l[[length(fisher.max.abs.l) + 1]] =
      if (all(!is.nan(fisher.2))) max(abs(fisher.2)) else Inf;
  }
  #print(fisher.max.abs.l);
  fisher.min = which.min(as.numeric(fisher.max.abs.l));
  print(sprintf("successful factor: %.3e", factors[[fisher.min]]));
  fisher.2 = fisher.l[[fisher.min]];
  }
  if (1) {
    fisher.2 = -hessian(likelihood, mle, h, NULL,
      d$reconstruction$diplotypes, d$sim$phenotypes, d$reconstruction$configCounts,
      d$sim$familySizes, clo, baseline
    );
  }
  fisher.2 = fisher.2 / (length(d$sim$familySizes));
  fisher.2
}
```

End of source code.

C.3. **Data simulations for a single observed locus.** All tables contain results for misspecifiation, *i.e.* simulation and estimation iterates all possible combinations for recessive, additive and dominant penetrance models.

Table C.1: Summary of parameter combinations simulated in the single locus case. Column A specifies whether samples were ascertained. n is the total number of individuals.

Table	Page	A	n	η_1	R_1	$p^\star$	β
C.2	123	-	300	0.1	0.0	0.1	2.0
C.3	125	-	300	0.1	0.5	0.1	2.0
C.4	127	-	300	0.1	0.9	0.1	2.0
C.5	129	-	300	0.2	0.0	0.2	2.0
C.6	131	-	300	0.2	0.5	0.2	2.0
C.7	133	-	300	0.2	0.9	0.2	2.0
C.8	135	-	300	0.1	0.5	0.2	2.0
C.9	137	-	300	0.2	0.5	0.1	2.0
C.10	139	-	300	0.3	0.9	0.3	2.0
C.11	141	-	1200	0.1	0.0	0.1	2.0
C.12	143	-	1200	0.1	0.5	0.1	2.0
C.13	145	-	1200	0.1	0.9	0.1	2.0
C.14	147	-	1200	0.2	0.0	0.2	2.0
C.15	149	-	1200	0.2	0.5	0.2	2.0
C.16	151	-	1200	0.2	0.9	0.2	2.0
C.17	153	-	1200	0.1	0.5	0.2	2.0
C.18	155	-	1200	0.2	0.5	0.1	2.0
C.19	157	-	1200	0.3	0.9	0.3	2.0
C.20	159	+	1200	0.1	0.0	0.1	2.0
C.21	160	+	1200	0.1	0.5	0.1	2.0
C.22	161	+	1200	0.1	0.9	0.1	2.0
C.23	162	+	1200	0.2	0.0	0.2	2.0
C.24	163	+	1200	0.2	0.5	0.2	2.0
C.25	164	+	1200	0.2	0.9	0.2	2.0
C.26	165	+	1200	0.1	0.5	0.2	2.0
C.27	166	+	1200	0.2	0.5	0.1	2.0
C.28	167	+	1200	0.3	0.9	0.3	2.0

Table C.2: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.0$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.19 (0.38; 0.18)	0.13 (0.16; 0.03)	1.14 (3.60; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.20 (0.38; 0.19)	0.16 (0.18; 0.04)	-3.10 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.05 (0.27; 0.08)	0.21 (0.22; 0.06)	2.26 (2.24; 5.10)
FS4	100	0.10 (0.02; 0.00)	0.14 (0.36; 0.15)	0.15 (0.19; 0.04)	1.36 (2.77; 8.06)
FS5	100	0.10 (0.02; 0.00)	0.13 (0.36; 0.15)	0.17 (0.21; 0.05)	1.20 (2.57; 7.27)
FS6	75	0.10 (0.02; 0.00)	0.03 (0.28; 0.08)	0.21 (0.22; 0.06)	2.17 (2.50; 6.28)
	simulation: dom; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.19 (0.37; 0.17)	0.24 (0.17; 0.05)	-6.55 (6.57; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.17 (0.36; 0.16)	0.24 (0.18; 0.05)	-10.03 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.33 (0.42; 0.29)	0.15 (0.16; 0.03)	-1.77 (6.91; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.24 (0.40; 0.22)	0.21 (0.19; 0.05)	-4.99 (6.65; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.27 (0.40; 0.23)	0.21 (0.18; 0.04)	-5.09 (6.44; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.33 (0.45; 0.31)	0.17 (0.18; 0.04)	-2.83 ($\geq$ 10; $\geq$ 10)
	simulation: dom; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.15 (0.36; 0.15)	0.14 (0.18; 0.03)	1.61 (3.59; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.16 (0.39; 0.18)	0.17 (0.20; 0.04)	0.69 (6.56; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.03 (0.27; 0.08)	0.19 (0.21; 0.05)	2.46 (2.89; 8.57)
FS4	100	0.10 (0.02; 0.00)	0.11 (0.33; 0.12)	0.15 (0.19; 0.04)	1.72 (2.68; 7.27)
FS5	100	0.10 (0.02; 0.00)	0.11 (0.33; 0.12)	0.15 (0.19; 0.04)	1.50 (2.96; 9.02)
FS6	75	0.10 (0.02; 0.00)	0.05 (0.26; 0.07)	0.18 (0.21; 0.05)	2.44 (2.11; 4.65)
	simulation: rec; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.10 (0.38; 0.16)	0.30 (0.26; 0.11)	3.36 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.10 (0.42; 0.18)	0.33 (0.29; 0.14)	2.45 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	-0.02 (0.28; 0.08)	0.38 (0.27; 0.15)	7.56 (6.60; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.04 (0.35; 0.12)	0.34 (0.27; 0.13)	4.27 (8.15; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.03 (0.36; 0.13)	0.36 (0.28; 0.15)	3.57 (5.68; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	-0.02 (0.29; 0.08)	0.41 (0.25; 0.16)	6.05 (7.04; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.27 (0.42; 0.25)	0.45 (0.26; 0.19)	-15.67 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.28 (0.42; 0.26)	0.45 (0.26; 0.19)	-17.46 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.25 (0.44; 0.26)	0.32 (0.25; 0.11)	-3.73 ($\geq$ 10; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.34 (0.43; 0.30)	0.37 (0.25; 0.14)	-13.03 ($\geq$ 10; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.35 (0.43; 0.31)	0.36 (0.25; 0.13)	-11.93 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.23 (0.44; 0.25)	0.33 (0.25; 0.12)	-3.65 ($\geq$ 10; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.07 (0.35; 0.13)	0.30 (0.26; 0.11)	4.96 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.05 (0.39; 0.16)	0.33 (0.29; 0.14)	4.10 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	-0.03 (0.27; 0.07)	0.36 (0.25; 0.13)	8.67 (9.25; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.02 (0.32; 0.10)	0.33 (0.26; 0.12)	5.13 (7.90; $\geq$ 10)

FS5	100	0.10 (0.02; 0.00)	0.02 (0.33; 0.11)	0.34 (0.27; 0.13)	4.74 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	-0.01 (0.26; 0.07)	0.37 (0.22; 0.12)	7.20 (6.92; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.17 (0.40; 0.19)	0.17 (0.23; 0.06)	1.28 (3.30; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.21 (0.42; 0.22)	0.21 (0.26; 0.08)	0.45 (2.55; 8.93)
FS3	100	0.10 (0.01; 0.00)	0.01 (0.32; 0.10)	0.29 (0.31; 0.13)	1.86 (2.94; 8.67)
FS4	100	0.10 (0.02; 0.00)	0.11 (0.37; 0.15)	0.19 (0.24; 0.07)	1.47 (4.32; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.14 (0.37; 0.16)	0.20 (0.25; 0.07)	0.94 (3.21; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.00 (0.30; 0.09)	0.26 (0.28; 0.11)	1.84 (3.09; 9.58)
	simulation: add; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.19 (0.39; 0.19)	0.25 (0.19; 0.06)	-6.34 (6.60; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.16 (0.39; 0.18)	0.26 (0.20; 0.07)	-6.49 (6.65; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.34 (0.45; 0.32)	0.17 (0.19; 0.04)	-1.44 (5.35; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.30 (0.42; 0.27)	0.22 (0.20; 0.05)	-4.93 (8.64; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.29 (0.43; 0.27)	0.22 (0.20; 0.05)	-4.78 (7.43; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.32 (0.46; 0.32)	0.20 (0.22; 0.06)	-2.32 (8.61; $\geq$ 10)
	simulation: add; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.14 (0.38; 0.16)	0.17 (0.23; 0.06)	1.57 (3.68; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.16 (0.40; 0.18)	0.21 (0.26; 0.08)	0.74 (3.67; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.01 (0.29; 0.08)	0.25 (0.29; 0.11)	2.32 (2.69; 7.33)
FS4	100	0.10 (0.02; 0.00)	0.07 (0.31; 0.10)	0.18 (0.24; 0.06)	1.56 (3.06; 9.56)
FS5	100	0.10 (0.02; 0.00)	0.10 (0.33; 0.12)	0.18 (0.24; 0.06)	1.44 (2.15; 4.92)
FS6	75	0.10 (0.02; 0.00)	0.02 (0.26; 0.07)	0.22 (0.24; 0.07)	2.27 (3.48; $\geq$ 10)

Table C.3: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.5$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.68 (0.29; 0.12)	0.09 (0.08; 0.01)	1.60 (1.91; 3.79)
FS2	150	0.10 (0.01; 0.00)	0.72 (0.27; 0.12)	0.09 (0.07; 0.01)	1.34 (1.31; 2.15)
FS3	100	0.10 (0.01; 0.00)	0.52 (0.27; 0.07)	0.13 (0.12; 0.01)	2.14 (7.28; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.62 (0.29; 0.10)	0.10 (0.09; 0.01)	1.77 (1.55; 2.45)
FS5	100	0.10 (0.02; 0.00)	0.64 (0.28; 0.10)	0.10 (0.09; 0.01)	1.67 (1.26; 1.68)
FS6	75	0.10 (0.02; 0.00)	0.53 (0.27; 0.07)	0.13 (0.11; 0.01)	2.20 (2.40; 5.80)
	simulation: dom; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.26 (0.44; 0.26)	0.26 (0.21; 0.07)	-4.36 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.21 (0.45; 0.28)	0.28 (0.23; 0.08)	-4.00 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.48 (0.43; 0.18)	0.15 (0.16; 0.03)	-1.27 ($\geq$ 10; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.35 (0.45; 0.22)	0.21 (0.20; 0.05)	-3.69 ($\geq$ 10; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.36 (0.44; 0.22)	0.22 (0.21; 0.06)	-3.89 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.47 (0.44; 0.19)	0.16 (0.19; 0.04)	-1.44 (8.41; $\geq$ 10)
	simulation: dom; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.66 (0.28; 0.10)	0.08 (0.07; 0.00)	1.92 (1.83; 3.35)
FS2	150	0.10 (0.01; 0.00)	0.69 (0.25; 0.10)	0.08 (0.07; 0.01)	1.66 (0.91; 0.95)
FS3	100	0.10 (0.01; 0.00)	0.51 (0.23; 0.05)	0.11 (0.09; 0.01)	2.64 (1.95; 4.21)
FS4	100	0.10 (0.02; 0.00)	0.62 (0.26; 0.08)	0.08 (0.07; 0.01)	2.01 (1.48; 2.18)
FS5	100	0.10 (0.02; 0.00)	0.61 (0.26; 0.08)	0.09 (0.08; 0.01)	1.92 (1.73; 2.99)
FS6	75	0.10 (0.02; 0.00)	0.52 (0.23; 0.05)	0.11 (0.08; 0.01)	2.51 (1.70; 3.14)
	simulation: rec; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.48 (0.22; 0.05)	0.23 (0.13; 0.03)	4.47 (8.16; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.50 (0.21; 0.04)	0.23 (0.13; 0.03)	4.38 (6.58; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.37 (0.17; 0.05)	0.31 (0.15; 0.07)	6.86 (5.46; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.45 (0.21; 0.04)	0.26 (0.14; 0.04)	4.34 (8.30; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.46 (0.19; 0.04)	0.27 (0.13; 0.05)	3.86 (5.42; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.37 (0.17; 0.05)	0.33 (0.15; 0.08)	5.37 (5.21; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.45 (0.45; 0.20)	0.36 (0.27; 0.14)	-13.43 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.41 (0.46; 0.22)	0.37 (0.27; 0.15)	-13.62 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.56 (0.38; 0.15)	0.25 (0.21; 0.07)	-8.22 ($\geq$ 10; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.49 (0.42; 0.18)	0.31 (0.24; 0.10)	-11.34 ($\geq$ 10; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.46 (0.44; 0.19)	0.33 (0.26; 0.12)	-10.63 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.53 (0.37; 0.14)	0.26 (0.20; 0.06)	-9.99 ($\geq$ 10; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.51 (0.19; 0.04)	0.21 (0.10; 0.02)	6.30 (8.48; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.52 (0.19; 0.04)	0.22 (0.12; 0.03)	5.28 (6.19; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.40 (0.16; 0.04)	0.28 (0.11; 0.05)	8.34 (5.90; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.49 (0.17; 0.03)	0.24 (0.10; 0.03)	5.42 (5.60; $\geq$ 10)

FS5	100	0.10 (0.02; 0.00)	0.49 (0.18; 0.03)	0.25 (0.12; 0.04)	4.88 (6.83; ≥ 10)
FS6	75	0.10 (0.02; 0.00)	0.41 (0.15; 0.03)	0.31 (0.11; 0.06)	6.47 (5.63; ≥ 10)
	simulation: add; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.66 (0.30; 0.12)	0.10 (0.10; 0.01)	1.54 (1.67; 2.99)
FS2	150	0.10 (0.01; 0.00)	0.72 (0.27; 0.12)	0.10 (0.10; 0.01)	1.28 (0.98; 1.47)
FS3	100	0.10 (0.01; 0.00)	0.49 (0.25; 0.06)	0.15 (0.15; 0.03)	2.11 (1.56; 2.46)
FS4	100	0.10 (0.02; 0.00)	0.61 (0.28; 0.09)	0.11 (0.11; 0.01)	1.67 (1.84; 3.48)
FS5	100	0.10 (0.02; 0.00)	0.64 (0.28; 0.10)	0.11 (0.11; 0.01)	1.54 (1.18; 1.60)
FS6	75	0.10 (0.02; 0.00)	0.50 (0.26; 0.07)	0.15 (0.14; 0.02)	2.06 (1.80; 3.25)
	simulation: add; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.29 (0.47; 0.27)	0.27 (0.24; 0.09)	-3.61 (6.65; ≥ 10)
FS2	150	0.10 (0.01; 0.00)	0.24 (0.49; 0.31)	0.30 (0.26; 0.11)	-3.48 (8.10; ≥ 10)
FS3	100	0.10 (0.01; 0.00)	0.51 (0.45; 0.20)	0.17 (0.20; 0.04)	-0.74 (9.71; ≥ 10)
FS4	100	0.10 (0.02; 0.00)	0.39 (0.49; 0.26)	0.24 (0.23; 0.07)	-2.80 (6.58; ≥ 10)
FS5	100	0.10 (0.02; 0.00)	0.39 (0.48; 0.24)	0.22 (0.22; 0.06)	-2.94 (6.16; ≥ 10)
FS6	75	0.10 (0.02; 0.00)	0.49 (0.47; 0.22)	0.18 (0.21; 0.05)	-3.70 (≥ 10; ≥ 10)
	simulation: add; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.69 (0.29; 0.12)	0.09 (0.08; 0.01)	1.72 (1.71; 3.02)
FS2	150	0.10 (0.01; 0.00)	0.73 (0.26; 0.12)	0.10 (0.09; 0.01)	1.47 (0.87; 1.03)
FS3	100	0.10 (0.01; 0.00)	0.52 (0.24; 0.06)	0.13 (0.12; 0.01)	2.35 (1.81; 3.40)
FS4	100	0.10 (0.02; 0.00)	0.63 (0.28; 0.09)	0.10 (0.08; 0.01)	1.87 (1.73; 3.00)
FS5	100	0.10 (0.02; 0.00)	0.63 (0.27; 0.09)	0.10 (0.09; 0.01)	1.81 (1.17; 1.41)
FS6	75	0.10 (0.02; 0.00)	0.52 (0.23; 0.06)	0.13 (0.10; 0.01)	2.23 (1.55; 2.46)

Table C.4: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.9$	$p^\star = 0.1$	$\beta = 2.0$
simulation: dom; estimation: dom					
FS1	150	0.10 (0.01; 0.00)	0.92 (0.15; 0.02)	0.09 (0.02; 0.00)	1.51 (1.80; 3.50)
FS2	150	0.10 (0.01; 0.00)	0.93 (0.12; 0.02)	0.09 (0.02; 0.00)	1.52 (0.43; 0.42)
FS3	100	0.10 (0.02; 0.00)	0.81 (0.19; 0.04)	0.10 (0.04; 0.00)	2.25 (0.91; 0.89)
FS4	100	0.10 (0.02; 0.00)	0.89 (0.16; 0.03)	0.09 (0.03; 0.00)	1.78 (1.05; 1.16)
FS5	100	0.10 (0.02; 0.00)	0.91 (0.14; 0.02)	0.09 (0.03; 0.00)	1.72 (0.49; 0.32)
FS6	75	0.10 (0.02; 0.00)	0.83 (0.17; 0.03)	0.10 (0.04; 0.00)	2.22 (0.93; 0.91)
simulation: dom; estimation: rec					
FS1	150	0.10 (0.01; 0.00)	0.39 (0.47; 0.49)	0.19 (0.15; 0.03)	-11.10 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.40 (0.50; 0.50)	0.19 (0.17; 0.04)	-13.23 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.57 (0.41; 0.28)	0.12 (0.11; 0.01)	-1.28 ($\geq$ 10; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.53 (0.46; 0.35)	0.15 (0.14; 0.02)	-7.17 ($\geq$ 10; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.50 (0.46; 0.37)	0.15 (0.14; 0.02)	-6.54 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.59 (0.40; 0.25)	0.12 (0.13; 0.02)	-3.92 ($\geq$ 10; $\geq$ 10)
simulation: dom; estimation: add					
FS1	150	0.10 (0.01; 0.00)	0.85 (0.17; 0.03)	0.07 (0.02; 0.00)	1.91 (0.75; 0.57)
FS2	150	0.10 (0.01; 0.00)	0.87 (0.15; 0.02)	0.08 (0.02; 0.00)	1.82 (0.55; 0.34)
FS3	100	0.10 (0.01; 0.00)	0.75 (0.17; 0.05)	0.08 (0.03; 0.00)	2.60 (0.82; 1.04)
FS4	100	0.10 (0.02; 0.00)	0.85 (0.16; 0.03)	0.08 (0.03; 0.00)	1.99 (0.58; 0.33)
FS5	100	0.10 (0.02; 0.00)	0.84 (0.17; 0.03)	0.08 (0.03; 0.00)	2.00 (0.61; 0.38)
FS6	75	0.10 (0.02; 0.00)	0.79 (0.16; 0.04)	0.09 (0.03; 0.00)	2.43 (0.67; 0.64)
simulation: rec; estimation: dom					
FS1	150	0.10 (0.01; 0.00)	0.64 (0.20; 0.11)	0.18 (0.06; 0.01)	3.11 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.63 (0.19; 0.11)	0.19 (0.06; 0.01)	3.24 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.52 (0.18; 0.18)	0.24 (0.07; 0.02)	9.46 ($\geq$ 10; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.61 (0.19; 0.12)	0.20 (0.07; 0.02)	2.49 ($\geq$ 10; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.59 (0.17; 0.12)	0.22 (0.07; 0.02)	3.54 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.52 (0.18; 0.18)	0.25 (0.08; 0.03)	9.38 ($\geq$ 10; $\geq$ 10)
simulation: rec; estimation: rec					
FS1	150	0.10 (0.01; 0.00)	0.64 (0.42; 0.25)	0.27 (0.25; 0.09)	-66.47 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.65 (0.44; 0.26)	0.26 (0.24; 0.08)	-99.54 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.72 (0.34; 0.15)	0.19 (0.16; 0.03)	-13.10 ($\geq$ 10; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.65 (0.41; 0.23)	0.24 (0.22; 0.07)	-45.99 ($\geq$ 10; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.68 (0.40; 0.21)	0.23 (0.22; 0.07)	-47.16 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.74 (0.33; 0.14)	0.19 (0.17; 0.04)	-21.87 ($\geq$ 10; $\geq$ 10)
simulation: rec; estimation: add					
FS1	150	0.10 (0.01; 0.00)	0.63 (0.11; 0.08)	0.18 (0.04; 0.01)	9.60 (7.62; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.64 (0.10; 0.08)	0.19 (0.04; 0.01)	7.69 (6.36; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.54 (0.10; 0.14)	0.24 (0.05; 0.02)	14.10 ($\geq$ 10; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.61 (0.10; 0.10)	0.20 (0.05; 0.01)	10.07 ($\geq$ 10; $\geq$ 10)

FS5	100	0.10 (0.02; 0.00)	0.62 (0.11; 0.09)	0.21 (0.05; 0.01)	7.26 (8.60; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.54 (0.09; 0.14)	0.24 (0.05; 0.02)	13.18 ($\geq$ 10; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.88 (0.18; 0.03)	0.08 (0.03; 0.00)	1.54 (0.83; 0.90)
FS2	150	0.10 (0.01; 0.00)	0.89 (0.17; 0.03)	0.08 (0.03; 0.00)	1.52 (0.59; 0.57)
FS3	100	0.10 (0.02; 0.00)	0.74 (0.18; 0.06)	0.09 (0.05; 0.00)	2.28 (1.15; 1.41)
FS4	100	0.10 (0.02; 0.00)	0.86 (0.18; 0.03)	0.08 (0.03; 0.00)	1.72 (0.95; 0.97)
FS5	100	0.10 (0.02; 0.00)	0.86 (0.17; 0.03)	0.09 (0.03; 0.00)	1.73 (0.78; 0.69)
FS6	75	0.10 (0.02; 0.00)	0.76 (0.18; 0.05)	0.10 (0.06; 0.00)	2.17 (0.90; 0.84)
	simulation: add; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.49 (0.47; 0.39)	0.18 (0.14; 0.03)	-7.72 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.48 (0.49; 0.42)	0.18 (0.16; 0.03)	-8.06 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.68 (0.40; 0.21)	0.13 (0.09; 0.01)	-2.85 ($\geq$ 10; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.62 (0.45; 0.28)	0.15 (0.13; 0.02)	-5.87 ($\geq$ 10; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.61 (0.45; 0.29)	0.15 (0.11; 0.02)	-6.15 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.65 (0.42; 0.24)	0.14 (0.14; 0.02)	-28.34 ($\geq$ 10; $\geq$ 10)
	simulation: add; estimation: add				
FS1	150	0.09 (0.01; 0.00)	0.93 (0.14; 0.02)	0.09 (0.02; 0.00)	1.63 (0.70; 0.62)
FS2	150	0.10 (0.01; 0.00)	0.95 (0.10; 0.01)	0.09 (0.02; 0.00)	1.55 (0.40; 0.36)
FS3	100	0.10 (0.01; 0.00)	0.83 (0.16; 0.03)	0.10 (0.04; 0.00)	2.20 (0.88; 0.82)
FS4	100	0.10 (0.02; 0.00)	0.90 (0.15; 0.02)	0.09 (0.03; 0.00)	1.79 (0.86; 0.78)
FS5	100	0.10 (0.02; 0.00)	0.91 (0.15; 0.02)	0.09 (0.03; 0.00)	1.70 (1.34; 1.89)
FS6	75	0.10 (0.02; 0.00)	0.84 (0.16; 0.03)	0.10 (0.04; 0.00)	2.16 (0.75; 0.59)

Table C.5: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.0$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	150	0.20 (0.02; 0.00)	-0.00 (0.34; 0.12)	0.28 (0.23; 0.06)	1.52 (1.12; 1.49)
FS2	150	0.20 (0.02; 0.00)	0.01 (0.37; 0.14)	0.33 (0.23; 0.07)	1.23 (0.62; 0.98)
FS3	100	0.20 (0.02; 0.00)	-0.00 (0.32; 0.10)	0.30 (0.23; 0.06)	2.10 (1.20; 1.44)
FS4	100	0.20 (0.02; 0.00)	0.01 (0.33; 0.11)	0.28 (0.23; 0.06)	1.62 (0.97; 1.10)
FS5	100	0.20 (0.02; 0.00)	0.01 (0.34; 0.11)	0.32 (0.24; 0.07)	1.49 (0.65; 0.68)
FS6	75	0.20 (0.02; 0.00)	0.00 (0.31; 0.10)	0.31 (0.23; 0.06)	1.93 (0.69; 0.49)
	simulation: dom; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.18 (0.48; 0.27)	0.27 (0.19; 0.04)	-3.09 (5.76; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.18 (0.47; 0.25)	0.27 (0.20; 0.05)	-2.83 (4.26; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.21 (0.49; 0.28)	0.21 (0.16; 0.02)	0.10 (2.97; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.21 (0.47; 0.27)	0.23 (0.18; 0.03)	-2.22 (5.02; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.24 (0.48; 0.29)	0.24 (0.18; 0.03)	-1.93 (4.85; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.20 (0.48; 0.27)	0.22 (0.18; 0.03)	0.23 (2.36; 8.71)
	simulation: dom; estimation: add				
FS1	150	0.20 (0.02; 0.00)	-0.02 (0.28; 0.08)	0.23 (0.22; 0.05)	1.91 (0.95; 0.91)
FS2	150	0.20 (0.02; 0.00)	0.00 (0.34; 0.12)	0.30 (0.23; 0.06)	1.54 (0.66; 0.65)
FS3	100	0.20 (0.02; 0.00)	0.00 (0.25; 0.06)	0.24 (0.19; 0.04)	2.59 (1.26; 1.92)
FS4	100	0.20 (0.02; 0.00)	0.02 (0.25; 0.06)	0.22 (0.19; 0.04)	2.06 (0.84; 0.71)
FS5	100	0.20 (0.02; 0.00)	-0.01 (0.27; 0.08)	0.23 (0.18; 0.03)	1.95 (0.70; 0.49)
FS6	75	0.20 (0.02; 0.00)	-0.01 (0.23; 0.05)	0.22 (0.16; 0.03)	2.49 (0.80; 0.88)
	simulation: rec; estimation: dom				
FS1	150	0.20 (0.02; 0.00)	-0.06 (0.34; 0.12)	0.54 (0.30; 0.20)	3.50 (4.32; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	-0.07 (0.37; 0.14)	0.59 (0.30; 0.24)	2.57 (3.46; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	-0.05 (0.27; 0.08)	0.59 (0.26; 0.22)	3.98 (4.23; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	-0.04 (0.31; 0.10)	0.57 (0.26; 0.21)	2.67 (3.39; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	-0.06 (0.32; 0.11)	0.60 (0.26; 0.23)	2.22 (2.40; 5.80)
FS6	75	0.20 (0.02; 0.00)	-0.03 (0.28; 0.08)	0.62 (0.23; 0.23)	3.00 (3.42; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.30 (0.48; 0.32)	0.43 (0.22; 0.10)	-9.93 ($\geq$ 10; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.28 (0.50; 0.33)	0.43 (0.23; 0.11)	-10.42 ($\geq$ 10; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.15 (0.48; 0.25)	0.37 (0.26; 0.10)	-0.96 ($\geq$ 10; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.30 (0.50; 0.34)	0.38 (0.23; 0.09)	-5.77 ($\geq$ 10; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.30 (0.50; 0.34)	0.37 (0.23; 0.08)	-5.25 ($\geq$ 10; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.16 (0.49; 0.27)	0.38 (0.26; 0.10)	-0.39 ($\geq$ 10; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	150	0.20 (0.02; 0.00)	-0.04 (0.27; 0.07)	0.49 (0.27; 0.16)	4.52 (4.79; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	-0.05 (0.30; 0.09)	0.56 (0.28; 0.21)	2.85 (3.27; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	-0.02 (0.20; 0.04)	0.52 (0.22; 0.15)	5.60 (4.95; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	-0.01 (0.23; 0.05)	0.51 (0.23; 0.15)	3.34 (3.31; $\geq$ 10)

FS5	100	0.20 (0.02; 0.00)	-0.03 (0.24; 0.06)	0.52 (0.22; 0.15)	2.96 (3.06; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	-0.02 (0.21; 0.04)	0.52 (0.19; 0.14)	4.48 (4.30; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	150	0.20 (0.02; 0.00)	-0.03 (0.35; 0.13)	0.36 (0.31; 0.12)	1.26 (1.13; 1.82)
FS2	150	0.20 (0.02; 0.00)	-0.03 (0.39; 0.15)	0.44 (0.32; 0.16)	0.94 (0.70; 1.61)
FS3	100	0.20 (0.02; 0.00)	-0.04 (0.31; 0.10)	0.42 (0.31; 0.15)	1.57 (1.15; 1.50)
FS4	100	0.20 (0.02; 0.00)	-0.04 (0.33; 0.11)	0.36 (0.30; 0.12)	1.35 (1.24; 1.97)
FS5	100	0.20 (0.02; 0.00)	-0.03 (0.34; 0.12)	0.41 (0.30; 0.14)	1.11 (0.69; 1.28)
FS6	75	0.20 (0.02; 0.00)	-0.04 (0.30; 0.09)	0.42 (0.29; 0.13)	1.45 (0.84; 1.01)
	simulation: add; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.19 (0.51; 0.29)	0.32 (0.23; 0.07)	-2.03 (3.62; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.17 (0.52; 0.30)	0.32 (0.23; 0.07)	-2.07 (3.86; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.20 (0.51; 0.30)	0.26 (0.21; 0.05)	0.34 (2.31; 8.11)
FS4	100	0.20 (0.02; 0.00)	0.20 (0.53; 0.32)	0.28 (0.21; 0.05)	-1.35 (4.19; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.24 (0.52; 0.33)	0.28 (0.22; 0.05)	-1.30 (3.26; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.21 (0.50; 0.30)	0.26 (0.22; 0.05)	0.22 (2.62; $\geq$ 10)
	simulation: add; estimation: add				
FS1	150	0.20 (0.02; 0.00)	-0.03 (0.28; 0.08)	0.31 (0.30; 0.10)	1.62 (1.12; 1.41)
FS2	150	0.20 (0.02; 0.00)	-0.04 (0.35; 0.12)	0.43 (0.31; 0.15)	1.12 (0.73; 1.31)
FS3	100	0.20 (0.02; 0.00)	-0.02 (0.24; 0.06)	0.32 (0.27; 0.09)	2.04 (1.06; 1.13)
FS4	100	0.20 (0.02; 0.00)	0.01 (0.25; 0.06)	0.30 (0.26; 0.08)	1.65 (0.98; 1.07)
FS5	100	0.20 (0.02; 0.00)	-0.02 (0.26; 0.07)	0.31 (0.26; 0.08)	1.56 (0.84; 0.89)
FS6	75	0.20 (0.02; 0.00)	-0.01 (0.21; 0.04)	0.30 (0.23; 0.06)	1.99 (1.01; 1.03)

Table C.6: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.5$	$p^\star = 0.2$	$\beta = 2.0$
simulation: dom; estimation: dom					
FS1	150	0.20 (0.02; 0.00)	0.60 (0.27; 0.09)	0.18 (0.11; 0.01)	1.62 (0.86; 0.89)
FS2	150	0.20 (0.02; 0.00)	0.65 (0.25; 0.09)	0.20 (0.11; 0.01)	1.47 (0.55; 0.59)
FS3	100	0.20 (0.02; 0.00)	0.53 (0.23; 0.05)	0.22 (0.11; 0.01)	2.12 (0.73; 0.55)
FS4	100	0.20 (0.02; 0.00)	0.59 (0.25; 0.07)	0.19 (0.11; 0.01)	1.69 (0.57; 0.42)
FS5	100	0.20 (0.02; 0.00)	0.61 (0.24; 0.07)	0.20 (0.10; 0.01)	1.63 (0.54; 0.43)
FS6	75	0.20 (0.02; 0.00)	0.52 (0.23; 0.05)	0.23 (0.12; 0.01)	2.06 (0.59; 0.35)
simulation: dom; estimation: rec					
FS1	150	0.20 (0.02; 0.00)	0.21 (0.58; 0.42)	0.28 (0.19; 0.04)	-1.23 (3.05; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.20 (0.57; 0.41)	0.30 (0.21; 0.05)	-1.34 (3.49; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.43 (0.47; 0.22)	0.20 (0.17; 0.03)	-0.42 ($\geq$ 10; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.36 (0.55; 0.32)	0.24 (0.18; 0.03)	-0.58 (2.54; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.38 (0.53; 0.30)	0.24 (0.19; 0.04)	-0.68 (3.20; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.45 (0.44; 0.20)	0.19 (0.17; 0.03)	0.63 (4.78; $\geq$ 10)
simulation: dom; estimation: add					
FS1	150	0.19 (0.02; 0.00)	0.55 (0.21; 0.05)	0.14 (0.07; 0.01)	2.08 (0.66; 0.45)
FS2	150	0.20 (0.02; 0.00)	0.60 (0.20; 0.05)	0.15 (0.07; 0.01)	1.90 (0.52; 0.28)
FS3	100	0.20 (0.02; 0.00)	0.48 (0.17; 0.03)	0.16 (0.08; 0.01)	2.69 (0.65; 0.90)
FS4	100	0.20 (0.02; 0.00)	0.54 (0.20; 0.04)	0.15 (0.07; 0.01)	2.16 (0.57; 0.35)
FS5	100	0.20 (0.02; 0.00)	0.55 (0.18; 0.04)	0.15 (0.08; 0.01)	2.14 (0.53; 0.30)
FS6	75	0.20 (0.02; 0.00)	0.50 (0.17; 0.03)	0.18 (0.08; 0.01)	2.56 (0.56; 0.63)
simulation: rec; estimation: dom					
FS1	150	0.20 (0.02; 0.00)	0.41 (0.18; 0.04)	0.41 (0.17; 0.07)	3.66 (5.64; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.44 (0.16; 0.03)	0.43 (0.16; 0.08)	2.69 (2.77; 8.16)
FS3	100	0.20 (0.02; 0.00)	0.37 (0.14; 0.04)	0.50 (0.15; 0.11)	3.82 (3.78; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.41 (0.16; 0.03)	0.44 (0.16; 0.09)	3.15 (3.66; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.41 (0.15; 0.03)	0.46 (0.14; 0.09)	2.55 (2.33; 5.74)
FS6	75	0.20 (0.02; 0.00)	0.37 (0.14; 0.04)	0.52 (0.14; 0.12)	3.26 (3.14; $\geq$ 10)
simulation: rec; estimation: rec					
FS1	150	0.20 (0.02; 0.00)	0.61 (0.41; 0.19)	0.29 (0.22; 0.06)	-4.62 ($\geq$ 10; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.59 (0.44; 0.21)	0.30 (0.22; 0.06)	-4.89 ($\geq$ 10; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.57 (0.38; 0.15)	0.28 (0.19; 0.04)	0.29 ($\geq$ 10; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.65 (0.36; 0.15)	0.26 (0.18; 0.04)	-1.73 ($\geq$ 10; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.63 (0.39; 0.17)	0.27 (0.19; 0.04)	-2.28 ($\geq$ 10; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.56 (0.39; 0.15)	0.30 (0.21; 0.05)	0.47 ($\geq$ 10; $\geq$ 10)
simulation: rec; estimation: add					
FS1	150	0.19 (0.02; 0.00)	0.48 (0.13; 0.02)	0.36 (0.12; 0.04)	4.73 (4.45; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.51 (0.13; 0.02)	0.39 (0.11; 0.05)	3.46 (3.17; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.42 (0.11; 0.02)	0.42 (0.11; 0.06)	6.31 (5.24; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.47 (0.13; 0.02)	0.39 (0.10; 0.05)	3.79 (3.37; $\geq$ 10)

FS5	100	0.20 (0.02; 0.00)	0.48 (0.12; 0.02)	0.41 (0.11; 0.05)	3.27 (2.42; 7.47)
FS6	75	0.20 (0.02; 0.00)	0.42 (0.12; 0.02)	0.45 (0.11; 0.07)	4.65 (4.14; ≥ 10)
	simulation: add; estimation: dom				
FS1	150	0.20 (0.02; 0.00)	0.55 (0.27; 0.08)	0.21 (0.16; 0.03)	1.46 (0.88; 1.07)
FS2	150	0.20 (0.02; 0.00)	0.62 (0.24; 0.07)	0.24 (0.16; 0.03)	1.21 (0.63; 1.02)
FS3	100	0.20 (0.02; 0.00)	0.46 (0.21; 0.04)	0.26 (0.17; 0.03)	1.83 (0.87; 0.79)
FS4	100	0.20 (0.02; 0.00)	0.55 (0.25; 0.06)	0.23 (0.16; 0.03)	1.46 (0.94; 1.18)
FS5	100	0.20 (0.02; 0.00)	0.55 (0.24; 0.06)	0.25 (0.16; 0.03)	1.37 (0.66; 0.83)
FS6	75	0.20 (0.02; 0.00)	0.48 (0.20; 0.04)	0.29 (0.17; 0.04)	1.67 (0.85; 0.84)
	simulation: add; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.28 (0.59; 0.40)	0.32 (0.23; 0.07)	-1.09 (3.04; ≥ 10)
FS2	150	0.20 (0.02; 0.00)	0.25 (0.60; 0.42)	0.31 (0.21; 0.06)	-1.13 (2.90; ≥ 10)
FS3	100	0.20 (0.02; 0.00)	0.52 (0.48; 0.23)	0.24 (0.18; 0.04)	0.47 (2.12; 6.84)
FS4	100	0.20 (0.02; 0.00)	0.43 (0.57; 0.33)	0.27 (0.20; 0.04)	-0.49 (2.46; ≥ 10)
FS5	100	0.20 (0.02; 0.00)	0.38 (0.59; 0.36)	0.28 (0.19; 0.04)	-0.71 (3.24; ≥ 10)
FS6	75	0.20 (0.02; 0.00)	0.51 (0.48; 0.23)	0.24 (0.20; 0.04)	0.50 (1.69; 5.09)
	simulation: add; estimation: add				
FS1	150	0.19 (0.02; 0.00)	0.61 (0.22; 0.06)	0.18 (0.11; 0.01)	1.71 (0.76; 0.66)
FS2	150	0.20 (0.02; 0.00)	0.68 (0.20; 0.07)	0.21 (0.10; 0.01)	1.45 (0.56; 0.62)
FS3	100	0.20 (0.02; 0.00)	0.53 (0.18; 0.03)	0.23 (0.12; 0.01)	2.12 (0.86; 0.76)
FS4	100	0.20 (0.02; 0.00)	0.59 (0.21; 0.05)	0.19 (0.10; 0.01)	1.74 (0.63; 0.46)
FS5	100	0.20 (0.02; 0.00)	0.61 (0.19; 0.05)	0.21 (0.10; 0.01)	1.65 (0.59; 0.47)
FS6	75	0.20 (0.02; 0.00)	0.53 (0.18; 0.03)	0.23 (0.11; 0.01)	2.03 (0.64; 0.41)

Table C.7: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.9$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	150	0.19 (0.02; 0.00)	0.93 (0.13; 0.02)	0.18 (0.04; 0.00)	1.53 (0.34; 0.33)
FS2	150	0.20 (0.02; 0.00)	0.94 (0.10; 0.01)	0.19 (0.03; 0.00)	1.50 (0.30; 0.33)
FS3	100	0.20 (0.02; 0.00)	0.83 (0.17; 0.03)	0.19 (0.06; 0.00)	2.19 (0.52; 0.30)
FS4	100	0.20 (0.02; 0.00)	0.89 (0.14; 0.02)	0.18 (0.04; 0.00)	1.70 (0.35; 0.21)
FS5	100	0.20 (0.02; 0.00)	0.91 (0.13; 0.02)	0.18 (0.04; 0.00)	1.69 (0.34; 0.21)
FS6	75	0.20 (0.02; 0.00)	0.84 (0.15; 0.03)	0.19 (0.05; 0.00)	2.15 (0.41; 0.19)
	simulation: dom; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.30 (0.65; 0.78)	0.29 (0.19; 0.05)	-1.03 ($\geq$ 10; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.30 (0.63; 0.76)	0.29 (0.18; 0.04)	-0.72 (4.11; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.51 (0.41; 0.32)	0.17 (0.15; 0.02)	1.03 (2.56; 7.51)
FS4	100	0.20 (0.02; 0.00)	0.48 (0.56; 0.48)	0.23 (0.17; 0.03)	-0.07 (1.73; 7.26)
FS5	100	0.20 (0.02; 0.00)	0.46 (0.57; 0.52)	0.24 (0.17; 0.03)	-1.12 ($\geq$ 10; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.50 (0.41; 0.33)	0.18 (0.15; 0.02)	0.61 (4.30; $\geq$ 10)
	simulation: dom; estimation: add				
FS1	150	0.19 (0.02; 0.00)	0.78 (0.16; 0.04)	0.13 (0.04; 0.01)	2.13 (0.48; 0.24)
FS2	150	0.20 (0.02; 0.00)	0.81 (0.15; 0.03)	0.14 (0.04; 0.01)	2.07 (0.42; 0.18)
FS3	100	0.20 (0.02; 0.00)	0.71 (0.13; 0.05)	0.14 (0.04; 0.01)	2.85 (0.49; 0.96)
FS4	100	0.20 (0.02; 0.00)	0.77 (0.14; 0.04)	0.13 (0.04; 0.01)	2.27 (0.41; 0.24)
FS5	100	0.20 (0.02; 0.00)	0.78 (0.14; 0.03)	0.14 (0.04; 0.01)	2.25 (0.42; 0.24)
FS6	75	0.20 (0.02; 0.00)	0.74 (0.13; 0.04)	0.15 (0.04; 0.00)	2.71 (0.48; 0.74)
	simulation: rec; estimation: dom				
FS1	150	0.19 (0.02; 0.00)	0.55 (0.10; 0.13)	0.38 (0.09; 0.04)	3.26 (3.05; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.56 (0.09; 0.13)	0.39 (0.09; 0.04)	2.69 (2.25; 5.54)
FS3	100	0.20 (0.02; 0.00)	0.50 (0.08; 0.17)	0.44 (0.09; 0.07)	3.97 (3.39; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.54 (0.09; 0.14)	0.41 (0.09; 0.05)	2.53 ($\geq$ 10; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.54 (0.09; 0.14)	0.42 (0.08; 0.05)	2.53 (1.57; 2.73)
FS6	75	0.20 (0.02; 0.00)	0.50 (0.08; 0.17)	0.45 (0.09; 0.07)	3.68 (3.39; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.85 (0.30; 0.09)	0.23 (0.14; 0.02)	-0.17 (9.60; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.82 (0.32; 0.11)	0.22 (0.14; 0.02)	-0.26 ($\geq$ 10; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.77 (0.27; 0.09)	0.23 (0.12; 0.01)	2.66 (7.80; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.85 (0.24; 0.06)	0.21 (0.11; 0.01)	1.05 ($\geq$ 10; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.85 (0.24; 0.06)	0.21 (0.11; 0.01)	-0.89 ($\geq$ 10; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.79 (0.26; 0.08)	0.23 (0.12; 0.02)	2.20 (7.57; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	150	0.19 (0.02; 0.00)	0.67 (0.07; 0.06)	0.31 (0.05; 0.01)	5.03 (4.10; $\geq$ 10)
FS2	150	0.19 (0.02; 0.00)	0.68 (0.07; 0.05)	0.33 (0.05; 0.02)	3.74 (2.16; 7.70)
FS3	100	0.20 (0.02; 0.00)	0.61 (0.06; 0.09)	0.38 (0.05; 0.04)	5.46 (3.85; $\geq$ 10)
FS4	100	0.19 (0.02; 0.00)	0.66 (0.07; 0.06)	0.33 (0.06; 0.02)	4.57 (3.38; $\geq$ 10)

FS5	100	0.19 (0.02; 0.00)	0.66 (0.07; 0.06)	0.35 (0.06; 0.02)	3.78 (2.26; 8.29)
FS6	75	0.20 (0.02; 0.00)	0.60 (0.06; 0.10)	0.39 (0.06; 0.04)	5.00 (3.73; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	150	0.19 (0.02; 0.00)	0.83 (0.19; 0.04)	0.16 (0.06; 0.01)	1.50 (0.59; 0.59)
FS2	150	0.20 (0.02; 0.00)	0.86 (0.16; 0.03)	0.17 (0.06; 0.00)	1.42 (0.50; 0.58)
FS3	100	0.20 (0.02; 0.00)	0.68 (0.16; 0.07)	0.16 (0.08; 0.01)	2.25 (0.67; 0.51)
FS4	100	0.19 (0.02; 0.00)	0.80 (0.18; 0.04)	0.16 (0.07; 0.01)	1.65 (0.54; 0.42)
FS5	100	0.20 (0.02; 0.00)	0.81 (0.17; 0.04)	0.17 (0.07; 0.01)	1.63 (0.54; 0.43)
FS6	75	0.20 (0.02; 0.00)	0.71 (0.15; 0.06)	0.18 (0.09; 0.01)	2.07 (0.58; 0.34)
	simulation: add; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.54 (0.60; 0.48)	0.28 (0.16; 0.03)	-0.52 (3.34; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.50 (0.62; 0.54)	0.29 (0.18; 0.04)	-0.48 (2.28; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.76 (0.39; 0.17)	0.23 (0.12; 0.01)	0.63 (1.20; 3.32)
FS4	100	0.20 (0.02; 0.00)	0.66 (0.52; 0.33)	0.25 (0.14; 0.02)	-0.17 (4.30; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.65 (0.54; 0.35)	0.25 (0.15; 0.03)	-0.75 ($\geq$ 10; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.72 (0.42; 0.21)	0.23 (0.13; 0.02)	0.45 (1.28; 4.05)
	simulation: add; estimation: add				
FS1	150	0.19 (0.02; 0.00)	0.95 (0.09; 0.01)	0.18 (0.03; 0.00)	1.56 (0.28; 0.27)
FS2	150	0.19 (0.02; 0.00)	0.96 (0.07; 0.01)	0.19 (0.03; 0.00)	1.53 (0.23; 0.27)
FS3	100	0.20 (0.02; 0.00)	0.86 (0.12; 0.01)	0.20 (0.05; 0.00)	2.12 (0.43; 0.20)
FS4	100	0.19 (0.02; 0.00)	0.92 (0.11; 0.01)	0.18 (0.04; 0.00)	1.71 (0.33; 0.20)
FS5	100	0.19 (0.02; 0.00)	0.93 (0.10; 0.01)	0.19 (0.04; 0.00)	1.70 (0.29; 0.18)
FS6	75	0.20 (0.02; 0.00)	0.87 (0.12; 0.01)	0.20 (0.05; 0.00)	2.09 (0.37; 0.14)

Table C.8: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.5$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.55 (0.27; 0.08)	0.16 (0.11; 0.01)	1.72 (0.96; 1.00)
FS2	150	0.10 (0.01; 0.00)	0.59 (0.24; 0.07)	0.18 (0.11; 0.01)	1.50 (0.55; 0.55)
FS3	100	0.10 (0.01; 0.00)	0.46 (0.21; 0.05)	0.20 (0.11; 0.01)	2.25 (0.90; 0.87)
FS4	100	0.10 (0.02; 0.00)	0.52 (0.23; 0.05)	0.17 (0.11; 0.01)	1.81 (0.72; 0.56)
FS5	100	0.10 (0.02; 0.00)	0.54 (0.22; 0.05)	0.18 (0.11; 0.01)	1.72 (0.54; 0.37)
FS6	75	0.10 (0.02; 0.00)	0.48 (0.21; 0.04)	0.20 (0.11; 0.01)	2.14 (0.70; 0.50)
	simulation: dom; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.16 (0.46; 0.32)	0.29 (0.21; 0.05)	-2.00 (7.52; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.19 (0.48; 0.33)	0.29 (0.23; 0.06)	-1.70 (3.79; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.37 (0.43; 0.20)	0.19 (0.19; 0.04)	0.78 (3.60; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.32 (0.49; 0.27)	0.24 (0.21; 0.05)	-1.08 (3.55; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.31 (0.48; 0.27)	0.24 (0.21; 0.05)	-1.23 (4.22; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.42 (0.43; 0.19)	0.19 (0.18; 0.03)	-2.30 ($\geq$ 10; $\geq$ 10)
	simulation: dom; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.54 (0.20; 0.04)	0.12 (0.07; 0.01)	2.18 (0.65; 0.46)
FS2	150	0.10 (0.01; 0.00)	0.58 (0.19; 0.04)	0.14 (0.08; 0.01)	2.00 (0.51; 0.26)
FS3	100	0.10 (0.01; 0.00)	0.48 (0.15; 0.02)	0.16 (0.07; 0.01)	2.67 (0.67; 0.90)
FS4	100	0.10 (0.02; 0.00)	0.53 (0.18; 0.03)	0.14 (0.07; 0.01)	2.21 (0.70; 0.53)
FS5	100	0.10 (0.02; 0.00)	0.55 (0.18; 0.04)	0.15 (0.07; 0.01)	2.18 (0.56; 0.34)
FS6	75	0.10 (0.02; 0.00)	0.48 (0.16; 0.03)	0.17 (0.07; 0.01)	2.60 (0.77; 0.95)
	simulation: rec; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.36 (0.17; 0.05)	0.36 (0.16; 0.05)	4.19 (4.31; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.37 (0.15; 0.04)	0.38 (0.16; 0.06)	3.15 (3.11; $\geq$ 10)
FS3	100	0.10 (0.01; 0.00)	0.30 (0.13; 0.05)	0.45 (0.15; 0.09)	4.20 (3.79; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.34 (0.14; 0.04)	0.41 (0.16; 0.07)	3.42 (3.41; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.35 (0.13; 0.04)	0.42 (0.16; 0.07)	2.90 (2.56; 7.38)
FS6	75	0.10 (0.02; 0.00)	0.30 (0.12; 0.05)	0.48 (0.15; 0.10)	3.66 (3.42; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.39 (0.49; 0.25)	0.33 (0.25; 0.08)	-5.43 ($\geq$ 10; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.41 (0.49; 0.25)	0.32 (0.25; 0.08)	-4.07 ($\geq$ 10; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.42 (0.40; 0.17)	0.29 (0.23; 0.06)	2.33 (9.98; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.45 (0.44; 0.19)	0.29 (0.22; 0.06)	-3.36 ($\geq$ 10; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.44 (0.45; 0.21)	0.29 (0.23; 0.06)	-4.99 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.42 (0.40; 0.17)	0.30 (0.23; 0.06)	1.94 (8.87; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.43 (0.11; 0.02)	0.30 (0.09; 0.02)	5.68 (4.25; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.44 (0.11; 0.02)	0.33 (0.11; 0.03)	4.11 (3.29; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.37 (0.09; 0.03)	0.40 (0.09; 0.05)	5.79 (4.21; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.41 (0.10; 0.02)	0.35 (0.10; 0.03)	4.59 (3.64; $\geq$ 10)

FS5	100	0.10 (0.02; 0.00)	0.41 (0.11; 0.02)	0.37 (0.10; 0.04)	4.37 ($\geq$ 10; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.36 (0.10; 0.03)	0.42 (0.10; 0.06)	5.29 (4.43; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	150	0.10 (0.01; 0.00)	0.50 (0.24; 0.06)	0.18 (0.14; 0.02)	1.55 (0.99; 1.18)
FS2	150	0.10 (0.01; 0.00)	0.54 (0.22; 0.05)	0.22 (0.15; 0.02)	1.28 (0.59; 0.87)
FS3	100	0.10 (0.01; 0.00)	0.41 (0.18; 0.04)	0.25 (0.17; 0.03)	1.84 (0.91; 0.86)
FS4	100	0.10 (0.02; 0.00)	0.48 (0.21; 0.04)	0.21 (0.15; 0.02)	1.55 (1.13; 1.49)
FS5	100	0.10 (0.02; 0.00)	0.50 (0.21; 0.04)	0.22 (0.14; 0.02)	1.44 (0.74; 0.87)
FS6	75	0.10 (0.02; 0.00)	0.43 (0.19; 0.04)	0.26 (0.16; 0.03)	1.73 (0.67; 0.53)
	simulation: add; estimation: rec				
FS1	150	0.10 (0.01; 0.00)	0.21 (0.52; 0.36)	0.30 (0.24; 0.07)	-1.65 (3.94; $\geq$ 10)
FS2	150	0.10 (0.01; 0.00)	0.23 (0.54; 0.37)	0.30 (0.25; 0.07)	-1.49 (3.83; $\geq$ 10)
FS3	100	0.10 (0.02; 0.00)	0.43 (0.47; 0.22)	0.22 (0.22; 0.05)	0.62 (3.76; $\geq$ 10)
FS4	100	0.10 (0.02; 0.00)	0.34 (0.52; 0.30)	0.26 (0.23; 0.06)	-1.08 (5.37; $\geq$ 10)
FS5	100	0.10 (0.02; 0.00)	0.34 (0.53; 0.31)	0.26 (0.23; 0.06)	-1.22 (6.61; $\geq$ 10)
FS6	75	0.10 (0.02; 0.00)	0.42 (0.48; 0.24)	0.23 (0.24; 0.06)	0.17 (6.34; $\geq$ 10)
	simulation: add; estimation: add				
FS1	150	0.10 (0.01; 0.00)	0.57 (0.20; 0.04)	0.15 (0.08; 0.01)	1.82 (0.71; 0.53)
FS2	150	0.10 (0.01; 0.00)	0.61 (0.18; 0.05)	0.18 (0.10; 0.01)	1.61 (0.55; 0.45)
FS3	100	0.10 (0.02; 0.00)	0.48 (0.16; 0.03)	0.21 (0.11; 0.01)	2.21 (0.74; 0.59)
FS4	100	0.10 (0.02; 0.00)	0.54 (0.17; 0.03)	0.17 (0.09; 0.01)	1.85 (0.75; 0.59)
FS5	100	0.10 (0.02; 0.00)	0.56 (0.17; 0.03)	0.18 (0.09; 0.01)	1.75 (0.57; 0.38)
FS6	75	0.10 (0.02; 0.00)	0.49 (0.16; 0.02)	0.21 (0.10; 0.01)	2.14 (0.76; 0.60)

Table C.9: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.5$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	150	0.20 (0.02; 0.00)	0.61 (0.22; 0.06)	0.11 (0.06; 0.00)	1.40 (1.36; 2.20)
FS2	150	0.20 (0.02; 0.00)	0.63 (0.21; 0.06)	0.11 (0.07; 0.00)	1.32 (1.28; 2.12)
FS3	100	0.20 (0.02; 0.00)	0.54 (0.25; 0.06)	0.14 (0.11; 0.01)	2.24 (1.67; 2.84)
FS4	100	0.20 (0.02; 0.00)	0.60 (0.24; 0.07)	0.12 (0.08; 0.01)	1.57 (1.15; 1.49)
FS5	100	0.20 (0.02; 0.00)	0.61 (0.24; 0.07)	0.12 (0.08; 0.01)	1.48 (1.25; 1.83)
FS6	75	0.20 (0.02; 0.00)	0.56 (0.26; 0.07)	0.15 (0.11; 0.01)	1.99 (1.10; 1.20)
	simulation: dom; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.16 (0.43; 0.30)	0.24 (0.17; 0.05)	-6.38 (8.82; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.14 (0.43; 0.32)	0.25 (0.17; 0.05)	-5.50 (7.77; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.35 (0.41; 0.19)	0.17 (0.16; 0.03)	-1.10 (4.35; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.26 (0.43; 0.24)	0.21 (0.17; 0.04)	-4.28 (6.96; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.26 (0.42; 0.24)	0.21 (0.17; 0.04)	-4.33 (6.90; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.36 (0.41; 0.19)	0.18 (0.18; 0.04)	-0.90 (4.67; $\geq$ 10)
	simulation: dom; estimation: add				
FS1	150	0.20 (0.02; 0.00)	0.59 (0.21; 0.05)	0.10 (0.06; 0.00)	1.63 (1.20; 1.57)
FS2	150	0.20 (0.02; 0.00)	0.61 (0.19; 0.05)	0.10 (0.06; 0.00)	1.52 (0.75; 0.79)
FS3	100	0.20 (0.02; 0.00)	0.51 (0.22; 0.05)	0.12 (0.09; 0.01)	2.51 (1.61; 2.84)
FS4	100	0.20 (0.02; 0.00)	0.54 (0.22; 0.05)	0.10 (0.07; 0.01)	1.92 (1.40; 1.97)
FS5	100	0.20 (0.02; 0.00)	0.58 (0.21; 0.05)	0.10 (0.07; 0.01)	1.81 (0.99; 1.02)
FS6	75	0.20 (0.02; 0.00)	0.52 (0.20; 0.04)	0.12 (0.09; 0.01)	2.38 (1.56; 2.58)
	simulation: rec; estimation: dom				
FS1	150	0.20 (0.02; 0.00)	0.48 (0.24; 0.06)	0.27 (0.15; 0.05)	4.33 (9.12; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.51 (0.23; 0.05)	0.27 (0.14; 0.05)	4.10 (6.65; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.39 (0.18; 0.04)	0.33 (0.15; 0.07)	6.81 (5.25; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.48 (0.21; 0.05)	0.31 (0.15; 0.07)	4.23 (5.62; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.49 (0.20; 0.04)	0.31 (0.15; 0.07)	3.47 (4.91; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.41 (0.19; 0.04)	0.36 (0.16; 0.09)	4.99 (5.12; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.40 (0.40; 0.17)	0.36 (0.27; 0.14)	-9.43 ($\geq$ 10; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.42 (0.41; 0.18)	0.35 (0.26; 0.13)	-8.69 ($\geq$ 10; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.51 (0.38; 0.14)	0.26 (0.22; 0.07)	-2.33 ($\geq$ 10; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.49 (0.39; 0.16)	0.30 (0.23; 0.10)	-6.85 ($\geq$ 10; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.50 (0.40; 0.16)	0.32 (0.24; 0.10)	-8.90 ($\geq$ 10; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.51 (0.39; 0.15)	0.26 (0.21; 0.07)	-1.98 ($\geq$ 10; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	150	0.20 (0.02; 0.00)	0.52 (0.21; 0.04)	0.25 (0.13; 0.04)	5.58 (9.28; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.55 (0.19; 0.04)	0.27 (0.13; 0.05)	4.65 (6.91; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.43 (0.16; 0.03)	0.31 (0.13; 0.06)	8.04 (5.73; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.51 (0.19; 0.04)	0.28 (0.13; 0.05)	4.88 (5.84; $\geq$ 10)

FS5	100	0.20 (0.02; 0.00)	0.52 (0.18; 0.03)	0.29 (0.14; 0.06)	4.23 (4.49; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.45 (0.16; 0.03)	0.33 (0.13; 0.07)	6.04 (5.16; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	150	0.20 (0.02; 0.00)	0.60 (0.23; 0.06)	0.12 (0.09; 0.01)	1.31 (1.69; 3.35)
FS2	150	0.20 (0.02; 0.00)	0.63 (0.21; 0.06)	0.12 (0.10; 0.01)	1.18 (0.74; 1.21)
FS3	100	0.20 (0.02; 0.00)	0.53 (0.24; 0.06)	0.16 (0.14; 0.02)	2.00 (1.44; 2.07)
FS4	100	0.20 (0.02; 0.00)	0.58 (0.26; 0.07)	0.13 (0.12; 0.02)	1.50 (1.21; 1.72)
FS5	100	0.20 (0.02; 0.00)	0.60 (0.23; 0.06)	0.13 (0.11; 0.01)	1.44 (1.29; 1.99)
FS6	75	0.20 (0.02; 0.00)	0.54 (0.25; 0.06)	0.17 (0.15; 0.03)	1.78 (1.24; 1.58)
	simulation: add; estimation: rec				
FS1	150	0.20 (0.02; 0.00)	0.18 (0.42; 0.28)	0.24 (0.17; 0.05)	-6.74 ($\geq$ 10; $\geq$ 10)
FS2	150	0.20 (0.02; 0.00)	0.15 (0.44; 0.32)	0.27 (0.21; 0.07)	-5.71 (7.34; $\geq$ 10)
FS3	100	0.20 (0.02; 0.00)	0.42 (0.40; 0.17)	0.18 (0.17; 0.04)	-0.43 (3.73; $\geq$ 10)
FS4	100	0.20 (0.02; 0.00)	0.27 (0.45; 0.25)	0.23 (0.20; 0.06)	-3.81 (6.55; $\geq$ 10)
FS5	100	0.20 (0.02; 0.00)	0.28 (0.45; 0.25)	0.23 (0.19; 0.06)	-4.15 (8.10; $\geq$ 10)
FS6	75	0.20 (0.02; 0.00)	0.41 (0.42; 0.18)	0.19 (0.19; 0.04)	-0.70 (5.11; $\geq$ 10)
	simulation: add; estimation: add				
FS1	150	0.20 (0.02; 0.00)	0.63 (0.23; 0.07)	0.11 (0.08; 0.01)	1.43 (1.20; 1.78)
FS2	150	0.20 (0.02; 0.00)	0.66 (0.20; 0.06)	0.12 (0.08; 0.01)	1.28 (0.77; 1.11)
FS3	100	0.20 (0.02; 0.00)	0.55 (0.24; 0.06)	0.15 (0.12; 0.02)	2.13 (1.63; 2.68)
FS4	100	0.20 (0.02; 0.00)	0.59 (0.22; 0.06)	0.11 (0.08; 0.01)	1.67 (1.10; 1.32)
FS5	100	0.20 (0.02; 0.00)	0.61 (0.21; 0.06)	0.13 (0.10; 0.01)	1.55 (1.16; 1.54)
FS6	75	0.20 (0.02; 0.00)	0.55 (0.24; 0.06)	0.15 (0.12; 0.02)	2.05 (1.58; 2.49)

Table C.10: Parameter estimates for a total of N = 300 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.3$	$R_1 = 0.9$	$p^\star = 0.3$	$\beta = 2.0$
simulation: dom; estimation: dom					
FS1	150	0.29 (0.02; 0.00)	0.92 (0.12; 0.02)	0.28 (0.05; 0.00)	1.53 (0.30; 0.31)
FS2	150	0.30 (0.02; 0.00)	0.94 (0.09; 0.01)	0.29 (0.04; 0.00)	1.48 (0.22; 0.32)
FS3	100	0.30 (0.02; 0.00)	0.84 (0.15; 0.02)	0.29 (0.07; 0.00)	2.12 (0.37; 0.15)
FS4	100	0.30 (0.02; 0.00)	0.90 (0.13; 0.02)	0.28 (0.05; 0.00)	1.70 (0.30; 0.18)
FS5	100	0.30 (0.03; 0.00)	0.91 (0.12; 0.01)	0.28 (0.05; 0.00)	1.67 (0.26; 0.18)
FS6	75	0.30 (0.03; 0.00)	0.85 (0.13; 0.02)	0.29 (0.06; 0.00)	2.08 (0.30; 0.10)
simulation: dom; estimation: rec					
FS1	150	0.30 (0.02; 0.00)	0.69 (0.43; 0.23)	0.27 (0.13; 0.02)	0.32 (0.71; 3.33)
FS2	150	0.30 (0.02; 0.00)	0.69 (0.43; 0.23)	0.27 (0.13; 0.02)	0.28 (0.74; 3.51)
FS3	100	0.30 (0.02; 0.00)	0.57 (0.29; 0.19)	0.19 (0.12; 0.03)	1.29 (1.00; 1.51)
FS4	100	0.30 (0.03; 0.00)	0.71 (0.34; 0.15)	0.25 (0.12; 0.02)	0.53 (0.81; 2.83)
FS5	100	0.30 (0.03; 0.00)	0.68 (0.36; 0.18)	0.24 (0.13; 0.02)	0.56 (1.05; 3.17)
FS6	75	0.30 (0.03; 0.00)	0.58 (0.31; 0.20)	0.20 (0.13; 0.03)	1.18 (0.78; 1.29)
simulation: dom; estimation: add					
FS1	150	0.29 (0.02; 0.00)	0.74 (0.15; 0.05)	0.19 (0.06; 0.02)	2.30 (0.43; 0.28)
FS2	150	0.30 (0.02; 0.00)	0.75 (0.14; 0.04)	0.20 (0.06; 0.01)	2.26 (0.37; 0.20)
FS3	100	0.30 (0.02; 0.00)	0.68 (0.11; 0.06)	0.19 (0.05; 0.01)	3.05 (0.41; 1.27)
FS4	100	0.29 (0.03; 0.00)	0.73 (0.13; 0.05)	0.19 (0.05; 0.01)	2.43 (0.33; 0.30)
FS5	100	0.30 (0.03; 0.00)	0.74 (0.13; 0.04)	0.20 (0.05; 0.01)	2.45 (0.35; 0.32)
FS6	75	0.30 (0.03; 0.00)	0.72 (0.11; 0.04)	0.21 (0.05; 0.01)	2.89 (0.35; 0.92)
simulation: rec; estimation: dom					
FS1	150	0.29 (0.02; 0.00)	0.54 (0.11; 0.14)	0.50 (0.11; 0.05)	2.41 (1.52; 2.48)
FS2	150	0.30 (0.02; 0.00)	0.55 (0.09; 0.13)	0.53 (0.09; 0.06)	2.14 (1.01; 1.04)
FS3	100	0.30 (0.02; 0.00)	0.49 (0.09; 0.17)	0.57 (0.10; 0.08)	3.01 (2.09; 5.40)
FS4	100	0.29 (0.03; 0.00)	0.53 (0.09; 0.14)	0.54 (0.10; 0.07)	2.27 (1.19; 1.48)
FS5	100	0.30 (0.03; 0.00)	0.53 (0.09; 0.14)	0.55 (0.09; 0.07)	2.19 (0.68; 0.50)
FS6	75	0.30 (0.03; 0.00)	0.49 (0.08; 0.17)	0.59 (0.09; 0.09)	2.71 (1.61; 3.10)
simulation: rec; estimation: rec					
FS1	150	0.30 (0.02; 0.00)	0.91 (0.18; 0.03)	0.29 (0.06; 0.00)	1.12 ($\geq$ 10; $\geq$ 10)
FS2	150	0.30 (0.02; 0.00)	0.91 (0.18; 0.03)	0.28 (0.06; 0.00)	1.41 (3.69; $\geq$ 10)
FS3	100	0.30 (0.02; 0.00)	0.81 (0.20; 0.05)	0.30 (0.10; 0.01)	3.47 (5.90; $\geq$ 10)
FS4	100	0.30 (0.03; 0.00)	0.89 (0.18; 0.03)	0.28 (0.08; 0.01)	1.99 (3.16; 9.98)
FS5	100	0.30 (0.03; 0.00)	0.89 (0.18; 0.03)	0.28 (0.07; 0.01)	1.85 (1.55; 2.44)
FS6	75	0.30 (0.03; 0.00)	0.81 (0.21; 0.05)	0.31 (0.11; 0.01)	2.94 (4.48; $\geq$ 10)
simulation: rec; estimation: add					
FS1	150	0.28 (0.02; 0.00)	0.70 (0.07; 0.04)	0.41 (0.05; 0.02)	3.75 (2.27; 8.22)
FS2	150	0.29 (0.02; 0.00)	0.71 (0.05; 0.04)	0.43 (0.05; 0.02)	3.20 (1.47; 3.60)
FS3	100	0.30 (0.02; 0.00)	0.64 (0.06; 0.07)	0.48 (0.05; 0.04)	4.48 (2.76; $\geq$ 10)
FS4	100	0.28 (0.03; 0.00)	0.68 (0.06; 0.05)	0.44 (0.05; 0.02)	3.55 (1.53; 4.74)

FS5	100	0.29 (0.03; 0.00)	0.68 (0.06; 0.05)	0.45 (0.05; 0.03)	3.29 (0.97; 2.63)
FS6	75	0.30 (0.03; 0.00)	0.64 (0.06; 0.07)	0.49 (0.05; 0.04)	4.05 (1.95; 8.03)
	simulation: add; estimation: dom				
FS1	150	0.29 (0.02; 0.00)	0.78 (0.20; 0.05)	0.26 (0.10; 0.01)	1.40 (0.57; 0.69)
FS2	150	0.30 (0.02; 0.00)	0.81 (0.17; 0.04)	0.27 (0.10; 0.01)	1.32 (0.50; 0.71)
FS3	100	0.30 (0.02; 0.00)	0.63 (0.15; 0.09)	0.26 (0.14; 0.02)	2.07 (0.72; 0.52)
FS4	100	0.29 (0.03; 0.00)	0.76 (0.18; 0.05)	0.28 (0.11; 0.01)	1.45 (0.51; 0.57)
FS5	100	0.30 (0.03; 0.00)	0.76 (0.17; 0.05)	0.28 (0.11; 0.01)	1.45 (0.53; 0.58)
FS6	75	0.30 (0.03; 0.00)	0.66 (0.14; 0.08)	0.29 (0.13; 0.02)	1.86 (0.62; 0.41)
	simulation: add; estimation: rec				
FS1	150	0.30 (0.02; 0.00)	0.90 (0.26; 0.07)	0.30 (0.06; 0.00)	0.30 (0.46; 3.11)
FS2	150	0.30 (0.02; 0.00)	0.88 (0.32; 0.10)	0.30 (0.07; 0.01)	0.26 (0.68; 3.50)
FS3	100	0.30 (0.02; 0.00)	0.90 (0.19; 0.03)	0.30 (0.06; 0.00)	0.70 (0.45; 1.90)
FS4	100	0.30 (0.03; 0.00)	0.90 (0.27; 0.07)	0.30 (0.06; 0.00)	0.35 (0.58; 3.05)
FS5	100	0.30 (0.02; 0.00)	0.91 (0.22; 0.05)	0.30 (0.05; 0.00)	0.41 (0.42; 2.72)
FS6	75	0.30 (0.03; 0.00)	0.89 (0.22; 0.05)	0.30 (0.07; 0.00)	0.68 (0.37; 1.89)
	simulation: add; estimation: add				
FS1	150	0.28 (0.02; 0.00)	0.95 (0.07; 0.01)	0.28 (0.04; 0.00)	1.54 (0.21; 0.26)
FS2	150	0.29 (0.02; 0.00)	0.97 (0.05; 0.01)	0.29 (0.03; 0.00)	1.51 (0.17; 0.26)
FS3	100	0.30 (0.02; 0.00)	0.88 (0.10; 0.01)	0.30 (0.05; 0.00)	2.07 (0.32; 0.11)
FS4	100	0.28 (0.02; 0.00)	0.93 (0.09; 0.01)	0.28 (0.04; 0.00)	1.69 (0.24; 0.15)
FS5	100	0.29 (0.02; 0.00)	0.94 (0.07; 0.01)	0.29 (0.04; 0.00)	1.68 (0.22; 0.15)
FS6	75	0.30 (0.03; 0.00)	0.88 (0.09; 0.01)	0.30 (0.05; 0.00)	2.05 (0.29; 0.09)

Table C.11: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.0$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.05 (0.25; 0.06)	0.13 (0.17; 0.03)	1.72 (2.13; 4.62)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	-0.00 (0.19; 0.03)	0.19 (0.20; 0.05)	1.97 (0.89; 0.79)
FS4	400	0.10 (0.01; 0.00)	0.02 (0.20; 0.04)	0.13 (0.17; 0.03)	1.64 (0.85; 0.85)
FS5	400				
FS6	300				
	simulation: dom; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.05 (0.24; 0.06)	0.23 (0.16; 0.04)	-5.79 (4.49; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.30 (0.40; 0.25)	0.12 (0.13; 0.02)	-0.23 (4.01; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.11 (0.32; 0.12)	0.22 (0.19; 0.05)	-3.88 (4.37; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.29 (0.40; 0.24)	0.12 (0.13; 0.02)	-0.19 (2.73; $\geq$ 10)
	simulation: dom; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.03 (0.20; 0.04)	0.10 (0.15; 0.02)	2.14 (2.10; 4.41)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.00 (0.14; 0.02)	0.14 (0.15; 0.03)	2.25 (0.74; 0.62)
FS4	400	0.10 (0.01; 0.00)	0.01 (0.15; 0.02)	0.09 (0.13; 0.02)	2.05 (1.01; 1.02)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	-0.00 (0.12; 0.02)	0.12 (0.11; 0.01)	2.25 (0.71; 0.57)
	simulation: rec; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	-0.00 (0.27; 0.07)	0.29 (0.27; 0.11)	5.81 (5.41; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	-0.02 (0.16; 0.03)	0.33 (0.22; 0.10)	8.63 (5.66; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	-0.01 (0.21; 0.04)	0.33 (0.24; 0.11)	3.95 (4.12; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	-0.01 (0.15; 0.02)	0.37 (0.20; 0.11)	6.17 (5.89; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.13 (0.34; 0.13)	0.51 (0.19; 0.21)	-17.89 ($\geq$ 10; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.30 (0.42; 0.27)	0.24 (0.23; 0.07)	-2.61 ($\geq$ 10; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.26 (0.40; 0.23)	0.39 (0.21; 0.13)	-13.02 ($\geq$ 10; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.26 (0.42; 0.24)	0.26 (0.23; 0.08)	-0.87 ($\geq$ 10; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.00 (0.20; 0.04)	0.24 (0.23; 0.07)	7.24 (5.47; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	-0.01 (0.12; 0.02)	0.29 (0.16; 0.06)	9.39 (5.58; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	-0.00 (0.16; 0.02)	0.28 (0.19; 0.07)	4.67 (4.36; $\geq$ 10)

FS5	400				
FS6	300	0.10 (0.01; 0.00)	-0.01 (0.12; 0.01)	0.33 (0.15; 0.07)	7.31 (6.11; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.07 (0.27; 0.08)	0.14 (0.20; 0.04)	1.83 (7.92; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	-0.02 (0.20; 0.04)	0.26 (0.27; 0.10)	1.61 (0.88; 0.93)
FS4	400	0.10 (0.01; 0.00)	0.00 (0.20; 0.04)	0.17 (0.23; 0.06)	1.53 (1.19; 1.63)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	-0.00 (0.17; 0.03)	0.21 (0.21; 0.05)	1.63 (0.72; 0.65)
	simulation: add; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.05 (0.25; 0.06)	0.26 (0.20; 0.07)	-5.70 (4.76; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.30 (0.40; 0.25)	0.14 (0.16; 0.03)	-0.11 (3.31; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.12 (0.32; 0.12)	0.25 (0.24; 0.08)	-3.68 (4.36; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.30 (0.42; 0.27)	0.14 (0.17; 0.03)	-0.10 (2.73; $\geq$ 10)
	simulation: add; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.03 (0.19; 0.04)	0.12 (0.19; 0.03)	2.04 (2.14; 4.60)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	-0.01 (0.13; 0.02)	0.17 (0.19; 0.04)	2.00 (0.84; 0.71)
FS4	400	0.10 (0.01; 0.00)	0.01 (0.13; 0.02)	0.11 (0.16; 0.03)	1.86 (0.99; 1.00)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	-0.01 (0.12; 0.01)	0.15 (0.16; 0.03)	1.97 (0.84; 0.71)

Table C.12: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.5$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.76 (0.21; 0.11)	0.07 (0.03; 0.00)	1.38 (0.54; 0.68)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.53 (0.16; 0.03)	0.11 (0.05; 0.00)	2.06 (0.53; 0.29)
FS4	400	0.10 (0.01; 0.00)	0.65 (0.21; 0.07)	0.08 (0.04; 0.00)	1.65 (0.77; 0.72)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.53 (0.16; 0.02)	0.11 (0.05; 0.00)	2.01 (0.46; 0.21)
	simulation: dom; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	-0.01 (0.36; 0.40)	0.29 (0.21; 0.08)	-3.17 (5.94; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.47 (0.40; 0.16)	0.11 (0.12; 0.01)	0.54 (1.96; 5.97)
FS4	400	0.10 (0.01; 0.00)	0.12 (0.45; 0.34)	0.24 (0.21; 0.07)	-2.34 (4.20; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.45 (0.40; 0.16)	0.12 (0.15; 0.02)	0.46 (2.11; 6.82)
	simulation: dom; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.68 (0.20; 0.07)	0.06 (0.03; 0.00)	1.74 (0.57; 0.39)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.50 (0.14; 0.02)	0.09 (0.04; 0.00)	2.37 (0.52; 0.41)
FS4	400	0.10 (0.01; 0.00)	0.59 (0.17; 0.04)	0.07 (0.03; 0.00)	1.95 (0.46; 0.22)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.51 (0.12; 0.02)	0.09 (0.04; 0.00)	2.28 (0.43; 0.27)
	simulation: rec; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.47 (0.12; 0.01)	0.22 (0.08; 0.02)	4.36 (4.05; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.38 (0.08; 0.02)	0.31 (0.08; 0.05)	5.03 (4.15; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.45 (0.10; 0.01)	0.26 (0.08; 0.03)	3.29 (2.86; 9.87)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.39 (0.08; 0.02)	0.33 (0.08; 0.06)	3.85 (3.25; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.44 (0.48; 0.24)	0.31 (0.26; 0.11)	-8.27 ($\geq$ 10; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.62 (0.32; 0.11)	0.17 (0.15; 0.03)	0.08 (9.98; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.57 (0.43; 0.19)	0.22 (0.21; 0.06)	-9.89 ($\geq$ 10; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.61 (0.31; 0.11)	0.17 (0.14; 0.02)	-0.26 ($\geq$ 10; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.50 (0.10; 0.01)	0.20 (0.06; 0.01)	5.90 (4.58; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.40 (0.08; 0.02)	0.28 (0.07; 0.04)	7.09 (4.65; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.49 (0.09; 0.01)	0.24 (0.06; 0.02)	3.91 (2.90; $\geq$ 10)

FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.42 (0.07; 0.01)	0.31 (0.07; 0.05)	4.67 (3.72; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.73 (0.21; 0.10)	0.07 (0.04; 0.00)	1.37 (0.61; 0.77)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.50 (0.15; 0.02)	0.12 (0.07; 0.01)	1.87 (0.64; 0.42)
FS4	400	0.10 (0.01; 0.00)	0.63 (0.21; 0.06)	0.08 (0.05; 0.00)	1.58 (0.57; 0.51)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.50 (0.14; 0.02)	0.13 (0.07; 0.01)	1.80 (0.56; 0.35)
	simulation: add; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.00 (0.38; 0.40)	0.31 (0.26; 0.11)	-3.05 (5.86; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.50 (0.43; 0.19)	0.15 (0.18; 0.03)	0.43 (2.14; 7.05)
FS4	400	0.10 (0.01; 0.00)	0.17 (0.47; 0.33)	0.27 (0.26; 0.10)	-2.05 (4.42; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.51 (0.42; 0.18)	0.14 (0.16; 0.03)	0.34 (2.01; 6.79)
	simulation: add; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.74 (0.20; 0.10)	0.07 (0.03; 0.00)	1.50 (0.55; 0.56)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.52 (0.14; 0.02)	0.12 (0.06; 0.00)	2.01 (0.60; 0.36)
FS4	400	0.10 (0.01; 0.00)	0.64 (0.19; 0.05)	0.08 (0.04; 0.00)	1.72 (0.52; 0.35)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.52 (0.13; 0.02)	0.11 (0.05; 0.00)	1.99 (0.50; 0.25)

Table C.13: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.9$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.97 (0.07; 0.01)	0.09 (0.01; 0.00)	1.48 (0.22; 0.32)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.84 (0.13; 0.02)	0.09 (0.02; 0.00)	2.15 (0.36; 0.15)
FS4	400	0.10 (0.01; 0.00)	0.95 (0.08; 0.01)	0.09 (0.01; 0.00)	1.64 (0.23; 0.18)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.87 (0.11; 0.01)	0.10 (0.02; 0.00)	2.07 (0.28; 0.08)
	simulation: dom; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.17 (0.48; 0.77)	0.18 (0.12; 0.02)	-5.49 ($\geq$ 10; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.57 (0.36; 0.24)	0.09 (0.06; 0.00)	0.62 (1.60; 4.46)
FS4	400	0.10 (0.01; 0.00)	0.42 (0.52; 0.51)	0.15 (0.12; 0.02)	-1.86 (7.97; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.59 (0.36; 0.22)	0.09 (0.07; 0.01)	0.56 (0.95; 2.97)
	simulation: dom; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.88 (0.12; 0.02)	0.08 (0.02; 0.00)	1.77 (0.33; 0.16)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.77 (0.11; 0.03)	0.08 (0.02; 0.00)	2.53 (0.34; 0.40)
FS4	400	0.10 (0.01; 0.00)	0.86 (0.12; 0.02)	0.08 (0.02; 0.00)	1.96 (0.31; 0.10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.79 (0.10; 0.02)	0.08 (0.02; 0.00)	2.43 (0.29; 0.27)
	simulation: rec; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.56 (0.09; 0.13)	0.21 (0.05; 0.01)	4.59 (5.92; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.45 (0.06; 0.20)	0.25 (0.05; 0.03)	10.89 (8.36; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.53 (0.07; 0.14)	0.23 (0.05; 0.02)	5.15 (5.17; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.46 (0.06; 0.20)	0.26 (0.05; 0.03)	10.32 (9.22; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.90 (0.25; 0.06)	0.13 (0.12; 0.02)	-8.81 ($\geq$ 10; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.80 (0.24; 0.07)	0.12 (0.07; 0.01)	0.23 ($\geq$ 10; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.89 (0.22; 0.05)	0.11 (0.09; 0.01)	-4.96 ($\geq$ 10; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.79 (0.23; 0.06)	0.13 (0.10; 0.01)	1.59 ($\geq$ 10; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.61 (0.06; 0.09)	0.18 (0.02; 0.01)	8.35 (4.94; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.53 (0.04; 0.14)	0.24 (0.03; 0.02)	12.10 (8.45; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.60 (0.05; 0.09)	0.20 (0.03; 0.01)	7.80 (6.10; $\geq$ 10)

FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.53 (0.05; 0.14)	0.25 (0.03; 0.02)	10.09 (8.40; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.92 (0.13; 0.02)	0.08 (0.02; 0.00)	1.44 (0.36; 0.43)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.72 (0.11; 0.04)	0.08 (0.02; 0.00)	2.32 (0.40; 0.27)
FS4	400	0.10 (0.01; 0.00)	0.88 (0.13; 0.02)	0.08 (0.02; 0.00)	1.65 (0.37; 0.26)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.76 (0.11; 0.03)	0.08 (0.03; 0.00)	2.15 (0.37; 0.16)
	simulation: add; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.35 (0.53; 0.58)	0.17 (0.14; 0.02)	-8.41 ($\geq$ 10; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.76 (0.35; 0.14)	0.11 (0.04; 0.00)	0.30 (0.81; 3.53)
FS4	400	0.10 (0.01; 0.00)	0.53 (0.51; 0.40)	0.15 (0.13; 0.02)	-2.89 ($\geq$ 10; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.75 (0.35; 0.14)	0.11 (0.04; 0.00)	0.31 (1.07; 3.98)
	simulation: add; estimation: add				
FS1	600	0.09 (0.01; 0.00)	0.98 (0.05; 0.01)	0.09 (0.01; 0.00)	1.50 (0.18; 0.29)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.87 (0.10; 0.01)	0.10 (0.02; 0.00)	2.08 (0.32; 0.11)
FS4	400	0.10 (0.01; 0.00)	0.95 (0.07; 0.01)	0.09 (0.01; 0.00)	1.66 (0.23; 0.17)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.88 (0.09; 0.01)	0.10 (0.02; 0.00)	2.04 (0.28; 0.08)

Table C.14: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.0$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	-0.02 (0.27; 0.07)	0.28 (0.23; 0.06)	1.38 (0.58; 0.72)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	-0.02 (0.22; 0.05)	0.28 (0.20; 0.05)	1.94 (0.53; 0.28)
FS4	400	0.20 (0.01; 0.00)	-0.02 (0.24; 0.06)	0.26 (0.20; 0.05)	1.56 (0.48; 0.42)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	-0.02 (0.19; 0.04)	0.27 (0.17; 0.03)	1.92 (0.43; 0.19)
	simulation: dom; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.07 (0.38; 0.15)	0.27 (0.18; 0.04)	-2.18 (3.19; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.15 (0.43; 0.21)	0.21 (0.14; 0.02)	0.61 (0.85; 2.67)
FS4	400	0.20 (0.01; 0.00)	0.16 (0.44; 0.22)	0.22 (0.15; 0.02)	-1.34 (3.17; $\geq$ 10)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.16 (0.43; 0.21)	0.21 (0.16; 0.02)	0.68 (1.09; 2.93)
	simulation: dom; estimation: add				
FS1	600	0.20 (0.01; 0.00)	-0.01 (0.17; 0.03)	0.17 (0.15; 0.02)	1.92 (0.55; 0.31)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	-0.01 (0.12; 0.01)	0.18 (0.10; 0.01)	2.53 (0.45; 0.48)
FS4	400	0.20 (0.01; 0.00)	-0.00 (0.11; 0.01)	0.16 (0.10; 0.01)	2.11 (0.41; 0.18)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	-0.00 (0.09; 0.01)	0.17 (0.06; 0.01)	2.55 (0.34; 0.42)
	simulation: rec; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	-0.03 (0.21; 0.05)	0.56 (0.27; 0.20)	2.12 (2.11; 4.48)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	-0.01 (0.14; 0.02)	0.60 (0.21; 0.20)	2.61 (2.13; 4.90)
FS4	400	0.20 (0.01; 0.00)	-0.02 (0.17; 0.03)	0.57 (0.21; 0.18)	1.82 (0.81; 0.70)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	-0.01 (0.14; 0.02)	0.62 (0.16; 0.20)	2.10 (0.96; 0.93)
	simulation: rec; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.11 (0.39; 0.16)	0.45 (0.17; 0.09)	-8.78 (8.77; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.06 (0.41; 0.17)	0.35 (0.28; 0.10)	2.60 (4.24; $\geq$ 10)
FS4	400	0.20 (0.01; 0.00)	0.23 (0.43; 0.24)	0.35 (0.19; 0.06)	-5.36 (9.74; $\geq$ 10)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.04 (0.40; 0.17)	0.36 (0.28; 0.10)	2.45 (4.75; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	600	0.20 (0.01; 0.00)	-0.01 (0.13; 0.02)	0.45 (0.20; 0.10)	3.05 (2.43; 6.99)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	-0.00 (0.08; 0.01)	0.47 (0.13; 0.09)	3.98 (2.69; $\geq$ 10)
FS4	400	0.20 (0.01; 0.00)	-0.00 (0.09; 0.01)	0.45 (0.13; 0.08)	2.62 (0.99; 1.36)

FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.00 (0.08; 0.01)	0.49 (0.09; 0.09)	3.15 (1.62; 3.94)
	simulation: add; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	-0.03 (0.24; 0.06)	0.37 (0.30; 0.12)	1.06 (0.62; 1.27)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	-0.01 (0.16; 0.02)	0.39 (0.26; 0.10)	1.39 (0.61; 0.75)
FS4	400	0.20 (0.01; 0.00)	-0.01 (0.19; 0.03)	0.35 (0.26; 0.09)	1.14 (0.51; 1.01)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	-0.01 (0.14; 0.02)	0.37 (0.21; 0.08)	1.33 (0.48; 0.68)
	simulation: add; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.07 (0.37; 0.14)	0.32 (0.23; 0.07)	-1.84 (3.11; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.12 (0.42; 0.19)	0.25 (0.20; 0.04)	0.61 (1.09; 3.10)
FS4	400	0.20 (0.01; 0.00)	0.18 (0.46; 0.24)	0.25 (0.18; 0.04)	-1.38 (2.90; $\geq$ 10)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.10 (0.43; 0.20)	0.24 (0.20; 0.04)	0.82 (7.02; $\geq$ 10)
	simulation: add; estimation: add				
FS1	600	0.20 (0.01; 0.00)	-0.01 (0.14; 0.02)	0.23 (0.21; 0.05)	1.54 (0.63; 0.61)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	-0.00 (0.09; 0.01)	0.23 (0.14; 0.02)	2.03 (0.57; 0.33)
FS4	400	0.20 (0.01; 0.00)	0.00 (0.09; 0.01)	0.20 (0.12; 0.01)	1.69 (0.50; 0.34)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	-0.00 (0.08; 0.01)	0.23 (0.09; 0.01)	1.98 (0.47; 0.22)

Table C.15: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.5$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	0.64 (0.18; 0.05)	0.17 (0.07; 0.01)	1.49 (0.41; 0.42)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.53 (0.15; 0.02)	0.22 (0.08; 0.01)	2.02 (0.40; 0.16)
FS4	400	0.20 (0.01; 0.00)	0.59 (0.17; 0.04)	0.19 (0.07; 0.01)	1.63 (0.36; 0.26)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.52 (0.14; 0.02)	0.21 (0.07; 0.01)	2.00 (0.32; 0.10)
	simulation: dom; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	-0.05 (0.55; 0.60)	0.33 (0.19; 0.06)	-0.62 (1.23; 8.39)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.47 (0.32; 0.10)	0.13 (0.11; 0.02)	1.05 (1.07; 2.05)
FS4	400	0.20 (0.01; 0.00)	0.31 (0.59; 0.38)	0.25 (0.17; 0.03)	-0.10 (0.98; 5.38)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.47 (0.31; 0.10)	0.13 (0.11; 0.02)	1.02 (0.91; 1.80)
	simulation: dom; estimation: add				
FS1	600	0.19 (0.01; 0.00)	0.53 (0.11; 0.01)	0.12 (0.04; 0.01)	2.07 (0.34; 0.12)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.46 (0.08; 0.01)	0.15 (0.04; 0.00)	2.62 (0.31; 0.48)
FS4	400	0.20 (0.01; 0.00)	0.50 (0.10; 0.01)	0.13 (0.04; 0.01)	2.19 (0.28; 0.11)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.47 (0.09; 0.01)	0.16 (0.04; 0.00)	2.56 (0.28; 0.39)
	simulation: rec; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	0.46 (0.10; 0.01)	0.43 (0.10; 0.06)	2.23 (1.55; 2.46)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.39 (0.08; 0.02)	0.52 (0.09; 0.11)	2.57 (1.29; 1.98)
FS4	400	0.20 (0.01; 0.00)	0.44 (0.08; 0.01)	0.47 (0.08; 0.08)	2.07 (0.55; 0.31)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.39 (0.07; 0.02)	0.54 (0.08; 0.12)	2.33 (0.67; 0.56)
	simulation: rec; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.64 (0.44; 0.22)	0.24 (0.18; 0.03)	-1.26 ($\geq$ 10; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.54 (0.27; 0.07)	0.26 (0.16; 0.03)	3.46 (4.47; $\geq$ 10)
FS4	400	0.20 (0.01; 0.00)	0.68 (0.34; 0.15)	0.20 (0.13; 0.02)	1.26 (4.28; $\geq$ 10)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.55 (0.26; 0.07)	0.26 (0.16; 0.03)	3.09 (4.05; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	600	0.19 (0.01; 0.00)	0.51 (0.08; 0.01)	0.37 (0.07; 0.03)	3.10 (1.82; 4.52)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.43 (0.06; 0.01)	0.44 (0.08; 0.06)	4.49 (3.31; $\geq$ 10)
FS4	400	0.20 (0.01; 0.00)	0.49 (0.07; 0.00)	0.41 (0.06; 0.05)	2.79 (0.68; 1.09)

FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.45 (0.06; 0.01)	0.47 (0.06; 0.07)	3.35 (1.90; 5.43)
	simulation: add; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	0.58 (0.18; 0.04)	0.19 (0.10; 0.01)	1.31 (0.52; 0.75)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.46 (0.12; 0.02)	0.26 (0.12; 0.02)	1.65 (0.53; 0.40)
FS4	400	0.20 (0.01; 0.00)	0.55 (0.15; 0.02)	0.22 (0.10; 0.01)	1.32 (0.42; 0.65)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.47 (0.11; 0.01)	0.27 (0.11; 0.02)	1.57 (0.47; 0.41)
	simulation: add; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.09 (0.61; 0.54)	0.31 (0.19; 0.05)	-0.57 (1.41; 8.60)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.67 (0.33; 0.14)	0.20 (0.12; 0.02)	0.77 (1.46; 3.63)
FS4	400	0.20 (0.01; 0.00)	0.48 (0.58; 0.33)	0.23 (0.14; 0.02)	-0.17 (1.10; 5.92)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.65 (0.35; 0.14)	0.20 (0.14; 0.02)	0.64 (0.66; 2.28)
	simulation: add; estimation: add				
FS1	600	0.19 (0.01; 0.00)	0.64 (0.15; 0.04)	0.18 (0.07; 0.01)	1.53 (0.43; 0.40)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.52 (0.11; 0.01)	0.22 (0.07; 0.01)	1.99 (0.45; 0.20)
FS4	400	0.19 (0.01; 0.00)	0.57 (0.12; 0.02)	0.18 (0.06; 0.00)	1.68 (0.37; 0.24)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.52 (0.10; 0.01)	0.22 (0.07; 0.00)	1.98 (0.39; 0.15)

Table C.16: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.9$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.19 (0.01; 0.00)	0.97 (0.06; 0.01)	0.18 (0.02; 0.00)	1.48 (0.15; 0.29)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.87 (0.11; 0.01)	0.19 (0.04; 0.00)	2.08 (0.24; 0.07)
FS4	400	0.19 (0.01; 0.00)	0.95 (0.08; 0.01)	0.18 (0.02; 0.00)	1.63 (0.16; 0.16)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.87 (0.09; 0.01)	0.19 (0.03; 0.00)	2.05 (0.21; 0.05)
	simulation: dom; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.13 (0.70; 1.08)	0.34 (0.19; 0.06)	-0.35 (0.68; 5.97)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.53 (0.30; 0.23)	0.11 (0.09; 0.02)	1.20 (0.89; 1.44)
FS4	400	0.20 (0.01; 0.00)	0.47 (0.63; 0.58)	0.26 (0.18; 0.04)	0.11 (0.61; 3.95)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.52 (0.29; 0.23)	0.11 (0.10; 0.02)	1.17 (0.92; 1.54)
	simulation: dom; estimation: add				
FS1	600	0.19 (0.01; 0.00)	0.78 (0.13; 0.03)	0.13 (0.04; 0.01)	2.11 (0.32; 0.12)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.70 (0.07; 0.04)	0.13 (0.02; 0.01)	2.85 (0.23; 0.78)
FS4	400	0.19 (0.01; 0.00)	0.77 (0.09; 0.03)	0.13 (0.03; 0.01)	2.25 (0.24; 0.12)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.74 (0.06; 0.03)	0.14 (0.02; 0.00)	2.71 (0.20; 0.54)
	simulation: rec; estimation: dom				
FS1	600	0.19 (0.01; 0.00)	0.58 (0.05; 0.11)	0.39 (0.05; 0.04)	2.34 (0.93; 0.99)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.52 (0.04; 0.15)	0.46 (0.04; 0.07)	2.80 (0.74; 1.18)
FS4	400	0.20 (0.01; 0.00)	0.56 (0.04; 0.12)	0.42 (0.04; 0.05)	2.35 (0.30; 0.21)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.52 (0.03; 0.15)	0.48 (0.04; 0.08)	2.64 (0.27; 0.48)
	simulation: rec; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.96 (0.07; 0.01)	0.19 (0.02; 0.00)	1.49 (0.90; 1.07)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.83 (0.18; 0.04)	0.19 (0.06; 0.00)	3.50 (3.91; $\geq$ 10)
FS4	400	0.20 (0.01; 0.00)	0.92 (0.12; 0.02)	0.18 (0.04; 0.00)	2.00 (1.71; 2.94)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.82 (0.17; 0.04)	0.19 (0.06; 0.00)	3.29 (3.35; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	600	0.19 (0.01; 0.00)	0.69 (0.04; 0.05)	0.32 (0.03; 0.01)	3.77 (1.99; 7.08)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.62 (0.03; 0.08)	0.39 (0.03; 0.04)	4.00 (1.11; 5.22)
FS4	400	0.19 (0.01; 0.00)	0.67 (0.03; 0.06)	0.34 (0.03; 0.02)	3.41 (0.91; 2.83)

FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.61 (0.03; 0.09)	0.40 (0.03; 0.04)	3.74 (0.65; 3.46)
	simulation: add; estimation: dom				
FS1	600	0.19 (0.01; 0.00)	0.86 (0.16; 0.03)	0.16 (0.05; 0.00)	1.45 (0.42; 0.48)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.64 (0.08; 0.07)	0.13 (0.04; 0.01)	2.38 (0.34; 0.27)
FS4	400	0.19 (0.01; 0.00)	0.80 (0.14; 0.03)	0.15 (0.05; 0.01)	1.68 (0.39; 0.26)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.70 (0.08; 0.05)	0.16 (0.05; 0.00)	2.12 (0.36; 0.15)
	simulation: add; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.69 (0.55; 0.35)	0.24 (0.12; 0.02)	0.04 (0.53; 4.13)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.92 (0.18; 0.03)	0.20 (0.04; 0.00)	0.56 (0.27; 2.15)
FS4	400	0.20 (0.01; 0.00)	0.85 (0.39; 0.15)	0.22 (0.10; 0.01)	0.25 (0.30; 3.17)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.90 (0.18; 0.03)	0.20 (0.05; 0.00)	0.58 (0.32; 2.11)
	simulation: add; estimation: add				
FS1	600	0.19 (0.01; 0.00)	0.98 (0.03; 0.01)	0.18 (0.01; 0.00)	1.51 (0.11; 0.25)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.89 (0.07; 0.00)	0.20 (0.03; 0.00)	2.04 (0.22; 0.05)
FS4	400	0.19 (0.01; 0.00)	0.96 (0.05; 0.01)	0.19 (0.02; 0.00)	1.64 (0.13; 0.14)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.89 (0.07; 0.00)	0.20 (0.03; 0.00)	2.03 (0.19; 0.04)

Table C.17: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.5$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.59 (0.15; 0.03)	0.15 (0.05; 0.01)	1.57 (0.38; 0.33)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.50 (0.11; 0.01)	0.20 (0.06; 0.00)	2.07 (0.35; 0.13)
FS4	400	0.10 (0.01; 0.00)	0.55 (0.13; 0.02)	0.17 (0.05; 0.00)	1.68 (0.30; 0.19)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.49 (0.11; 0.01)	0.20 (0.05; 0.00)	2.02 (0.28; 0.08)
	simulation: dom; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	-0.06 (0.39; 0.47)	0.34 (0.22; 0.07)	-0.96 (1.86; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.46 (0.33; 0.11)	0.12 (0.12; 0.02)	1.24 (1.61; 3.16)
FS4	400	0.10 (0.01; 0.00)	0.23 (0.52; 0.34)	0.26 (0.23; 0.06)	-0.30 (3.72; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.47 (0.33; 0.11)	0.12 (0.13; 0.02)	1.15 (1.36; 2.56)
	simulation: dom; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.55 (0.12; 0.02)	0.12 (0.04; 0.01)	2.07 (0.34; 0.12)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.47 (0.09; 0.01)	0.15 (0.04; 0.00)	2.63 (0.31; 0.49)
FS4	400	0.10 (0.01; 0.00)	0.51 (0.10; 0.01)	0.13 (0.04; 0.01)	2.18 (0.28; 0.11)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.48 (0.09; 0.01)	0.16 (0.03; 0.00)	2.55 (0.26; 0.38)
	simulation: rec; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.40 (0.07; 0.01)	0.36 (0.07; 0.03)	2.62 (1.65; 3.11)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.33 (0.05; 0.03)	0.47 (0.07; 0.08)	2.75 (0.90; 1.38)
FS4	400	0.10 (0.01; 0.00)	0.38 (0.06; 0.02)	0.41 (0.07; 0.05)	2.38 (0.60; 0.50)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.33 (0.04; 0.03)	0.49 (0.06; 0.09)	2.56 (0.36; 0.44)
	simulation: rec; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	0.49 (0.48; 0.23)	0.25 (0.21; 0.05)	-1.55 ($\geq$ 10; $\geq$ 10)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.47 (0.28; 0.08)	0.22 (0.17; 0.03)	4.24 (5.17; $\geq$ 10)
FS4	400	0.10 (0.01; 0.00)	0.56 (0.39; 0.16)	0.19 (0.16; 0.03)	1.13 (8.93; $\geq$ 10)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.49 (0.27; 0.07)	0.22 (0.15; 0.02)	3.67 (5.43; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.45 (0.05; 0.01)	0.31 (0.05; 0.02)	3.59 (1.74; 5.56)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.38 (0.04; 0.02)	0.41 (0.05; 0.04)	4.02 (1.92; 7.77)
FS4	400	0.10 (0.01; 0.00)	0.42 (0.05; 0.01)	0.36 (0.05; 0.03)	3.19 (0.89; 2.21)

FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.38 (0.04; 0.02)	0.43 (0.04; 0.05)	3.43 (0.95; 2.95)
	simulation: add; estimation: dom				
FS1	600	0.10 (0.01; 0.00)	0.54 (0.13; 0.02)	0.18 (0.07; 0.01)	1.33 (0.47; 0.67)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.43 (0.09; 0.01)	0.24 (0.09; 0.01)	1.68 (0.49; 0.34)
FS4	400	0.10 (0.01; 0.00)	0.50 (0.11; 0.01)	0.20 (0.08; 0.01)	1.39 (0.40; 0.53)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.45 (0.09; 0.01)	0.25 (0.08; 0.01)	1.60 (0.40; 0.31)
	simulation: add; estimation: rec				
FS1	600	0.10 (0.01; 0.00)	-0.03 (0.52; 0.55)	0.37 (0.26; 0.10)	-0.67 (1.58; 9.62)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.59 (0.34; 0.12)	0.17 (0.14; 0.02)	0.79 (1.08; 2.62)
FS4	400	0.10 (0.01; 0.00)	0.32 (0.57; 0.35)	0.27 (0.25; 0.07)	-0.19 (1.62; 7.40)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.55 (0.34; 0.12)	0.17 (0.14; 0.02)	0.79 (1.38; 3.36)
	simulation: add; estimation: add				
FS1	600	0.10 (0.01; 0.00)	0.59 (0.11; 0.02)	0.16 (0.05; 0.00)	1.64 (0.37; 0.27)
FS2	600				
FS3	400	0.10 (0.01; 0.00)	0.50 (0.08; 0.01)	0.21 (0.06; 0.00)	2.03 (0.39; 0.15)
FS4	400	0.10 (0.01; 0.00)	0.55 (0.10; 0.01)	0.17 (0.05; 0.00)	1.71 (0.33; 0.19)
FS5	400				
FS6	300	0.10 (0.01; 0.00)	0.50 (0.08; 0.01)	0.20 (0.05; 0.00)	2.03 (0.35; 0.13)

Table C.18: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.5$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	0.64 (0.12; 0.04)	0.10 (0.03; 0.00)	1.30 (0.41; 0.66)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.56 (0.19; 0.04)	0.13 (0.07; 0.01)	1.96 (0.55; 0.31)
FS4	400	0.20 (0.01; 0.00)	0.61 (0.16; 0.04)	0.10 (0.05; 0.00)	1.54 (0.43; 0.40)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.55 (0.16; 0.03)	0.12 (0.06; 0.00)	1.95 (0.45; 0.21)
	simulation: dom; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	-0.09 (0.31; 0.45)	0.24 (0.15; 0.04)	-4.85 (4.86; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.37 (0.35; 0.14)	0.13 (0.11; 0.01)	0.43 (1.48; 4.63)
FS4	400	0.20 (0.01; 0.00)	0.04 (0.44; 0.41)	0.24 (0.19; 0.06)	-2.45 (5.30; $\geq$ 10)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.38 (0.35; 0.13)	0.13 (0.13; 0.02)	0.26 (1.76; 6.13)
	simulation: dom; estimation: add				
FS1	600	0.20 (0.01; 0.00)	0.59 (0.12; 0.02)	0.08 (0.03; 0.00)	1.57 (0.45; 0.38)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.49 (0.13; 0.02)	0.09 (0.05; 0.00)	2.35 (0.49; 0.36)
FS4	400	0.20 (0.01; 0.00)	0.55 (0.13; 0.02)	0.08 (0.03; 0.00)	1.86 (0.43; 0.20)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.50 (0.12; 0.01)	0.10 (0.04; 0.00)	2.29 (0.41; 0.25)
	simulation: rec; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	0.49 (0.16; 0.02)	0.26 (0.12; 0.04)	4.20 (3.99; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.40 (0.12; 0.02)	0.34 (0.12; 0.07)	5.22 (4.20; $\geq$ 10)
FS4	400	0.20 (0.01; 0.00)	0.48 (0.13; 0.02)	0.31 (0.10; 0.05)	2.74 (2.23; 5.53)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.43 (0.11; 0.02)	0.37 (0.10; 0.08)	3.67 (3.27; $\geq$ 10)
	simulation: rec; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	0.23 (0.40; 0.24)	0.38 (0.27; 0.15)	-8.18 ($\geq$ 10; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.48 (0.31; 0.10)	0.22 (0.19; 0.05)	1.14 (6.18; $\geq$ 10)
FS4	400	0.20 (0.01; 0.00)	0.36 (0.39; 0.17)	0.29 (0.24; 0.09)	-4.83 ($\geq$ 10; $\geq$ 10)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.48 (0.32; 0.10)	0.21 (0.19; 0.05)	1.49 (7.51; $\geq$ 10)
	simulation: rec; estimation: add				
FS1	600	0.20 (0.01; 0.00)	0.51 (0.14; 0.02)	0.23 (0.09; 0.03)	5.38 (4.34; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.43 (0.11; 0.02)	0.30 (0.09; 0.05)	6.62 (4.38; $\geq$ 10)
FS4	400	0.20 (0.01; 0.00)	0.51 (0.11; 0.01)	0.28 (0.08; 0.04)	3.25 (2.33; 7.01)

FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.46 (0.10; 0.01)	0.33 (0.08; 0.06)	4.42 (3.66; $\geq$ 10)
	simulation: add; estimation: dom				
FS1	600	0.20 (0.01; 0.00)	0.65 (0.13; 0.04)	0.10 (0.04; 0.00)	1.18 (0.44; 0.86)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.51 (0.17; 0.03)	0.13 (0.09; 0.01)	1.87 (0.66; 0.46)
FS4	400	0.20 (0.01; 0.00)	0.60 (0.17; 0.04)	0.10 (0.06; 0.00)	1.43 (0.51; 0.58)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.51 (0.15; 0.02)	0.13 (0.08; 0.01)	1.79 (0.56; 0.36)
	simulation: add; estimation: rec				
FS1	600	0.20 (0.01; 0.00)	-0.07 (0.30; 0.42)	0.27 (0.20; 0.07)	-4.96 (8.29; $\geq$ 10)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.44 (0.36; 0.13)	0.15 (0.12; 0.02)	0.18 (1.36; 5.18)
FS4	400	0.20 (0.01; 0.00)	0.06 (0.40; 0.35)	0.25 (0.22; 0.07)	-2.56 (4.03; $\geq$ 10)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.43 (0.36; 0.14)	0.14 (0.14; 0.02)	0.31 (1.31; 4.58)
	simulation: add; estimation: add				
FS1	600	0.20 (0.01; 0.00)	0.65 (0.13; 0.04)	0.10 (0.03; 0.00)	1.33 (0.41; 0.62)
FS2	600				
FS3	400	0.20 (0.01; 0.00)	0.56 (0.17; 0.03)	0.13 (0.07; 0.01)	1.92 (0.61; 0.38)
FS4	400	0.20 (0.01; 0.00)	0.61 (0.16; 0.04)	0.10 (0.05; 0.00)	1.57 (0.47; 0.41)
FS5	400				
FS6	300	0.20 (0.01; 0.00)	0.53 (0.14; 0.02)	0.12 (0.06; 0.00)	1.98 (0.53; 0.28)

Table C.19: Parameter estimates for a total of N = 1200 individuals and unconditional sampling. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.3$	$R_1 = 0.9$	$p^\star = 0.3$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1	600	0.29 (0.01; 0.00)	0.97 (0.05; 0.01)	0.29 (0.03; 0.00)	1.47 (0.13; 0.30)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.87 (0.10; 0.01)	0.29 (0.05; 0.00)	2.06 (0.20; 0.04)
FS4	400	0.29 (0.01; 0.00)	0.95 (0.06; 0.01)	0.28 (0.03; 0.00)	1.63 (0.13; 0.15)
FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.88 (0.09; 0.01)	0.29 (0.04; 0.00)	2.04 (0.17; 0.03)
	simulation: dom; estimation: rec				
FS1	600	0.30 (0.01; 0.00)	0.75 (0.32; 0.12)	0.24 (0.11; 0.02)	0.41 (0.48; 2.74)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.55 (0.24; 0.18)	0.14 (0.10; 0.03)	1.40 (0.66; 0.79)
FS4	400	0.30 (0.01; 0.00)	0.74 (0.25; 0.09)	0.22 (0.10; 0.02)	0.62 (0.53; 2.19)
FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.54 (0.22; 0.18)	0.14 (0.09; 0.04)	1.40 (0.60; 0.72)
	simulation: dom; estimation: add				
FS1	600	0.29 (0.01; 0.00)	0.72 (0.12; 0.05)	0.18 (0.05; 0.02)	2.35 (0.31; 0.22)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.67 (0.07; 0.06)	0.19 (0.03; 0.01)	3.07 (0.23; 1.19)
FS4	400	0.29 (0.01; 0.00)	0.71 (0.07; 0.04)	0.18 (0.03; 0.02)	2.48 (0.18; 0.26)
FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.71 (0.06; 0.04)	0.20 (0.03; 0.01)	2.93 (0.19; 0.89)
	simulation: rec; estimation: dom				
FS1	600	0.29 (0.01; 0.00)	0.57 (0.05; 0.11)	0.54 (0.05; 0.06)	2.00 (0.28; 0.08)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.52 (0.04; 0.15)	0.59 (0.04; 0.09)	2.48 (0.23; 0.29)
FS4	400	0.30 (0.01; 0.00)	0.55 (0.04; 0.12)	0.56 (0.04; 0.07)	2.08 (0.20; 0.05)
FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.51 (0.04; 0.15)	0.61 (0.04; 0.10)	2.41 (0.18; 0.20)
	simulation: rec; estimation: rec				
FS1	600	0.30 (0.01; 0.00)	0.97 (0.05; 0.01)	0.29 (0.02; 0.00)	1.48 (0.69; 0.75)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.84 (0.15; 0.03)	0.28 (0.07; 0.01)	2.99 (2.82; 8.96)
FS4	400	0.30 (0.01; 0.00)	0.93 (0.10; 0.01)	0.28 (0.04; 0.00)	1.77 (0.90; 0.87)
FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.85 (0.14; 0.02)	0.28 (0.06; 0.00)	2.57 (1.85; 3.76)
	simulation: rec; estimation: add				
FS1	600	0.28 (0.01; 0.00)	0.72 (0.03; 0.03)	0.42 (0.02; 0.02)	3.20 (0.24; 1.51)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.65 (0.03; 0.06)	0.49 (0.02; 0.04)	3.75 (0.36; 3.21)
FS4	400	0.28 (0.01; 0.00)	0.70 (0.03; 0.04)	0.45 (0.02; 0.02)	3.24 (0.19; 1.57)

FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.65 (0.03; 0.06)	0.50 (0.02; 0.04)	3.63 (0.20; 2.71)
	simulation: add; estimation: dom				
FS1	600	0.29 (0.01; 0.00)	0.76 (0.17; 0.05)	0.23 (0.09; 0.01)	1.46 (0.47; 0.52)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.60 (0.09; 0.10)	0.22 (0.10; 0.02)	2.20 (0.52; 0.31)
FS4	400	0.29 (0.01; 0.00)	0.73 (0.14; 0.05)	0.25 (0.10; 0.01)	1.52 (0.43; 0.42)
FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.65 (0.09; 0.07)	0.27 (0.11; 0.01)	1.90 (0.48; 0.24)
	simulation: add; estimation: rec				
FS1	600	0.30 (0.01; 0.00)	0.96 (0.05; 0.01)	0.29 (0.02; 0.00)	0.34 (0.12; 2.76)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.96 (0.05; 0.01)	0.30 (0.02; 0.00)	0.67 (0.12; 1.78)
FS4	400	0.30 (0.01; 0.00)	0.96 (0.06; 0.01)	0.29 (0.02; 0.00)	0.44 (0.13; 2.44)
FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.96 (0.06; 0.01)	0.29 (0.03; 0.00)	0.68 (0.14; 1.75)
	simulation: add; estimation: add				
FS1	600	0.28 (0.01; 0.00)	0.98 (0.03; 0.01)	0.28 (0.02; 0.00)	1.51 (0.09; 0.24)
FS2	600				
FS3	400	0.30 (0.01; 0.00)	0.90 (0.06; 0.00)	0.30 (0.03; 0.00)	2.03 (0.17; 0.03)
FS4	400	0.29 (0.01; 0.00)	0.96 (0.05; 0.01)	0.28 (0.02; 0.00)	1.64 (0.10; 0.14)
FS5	400				
FS6	300	0.30 (0.01; 0.00)	0.90 (0.05; 0.00)	0.30 (0.03; 0.00)	2.02 (0.15; 0.02)

Table C.20: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.0$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1a	640	0.10 (0.01; 0.00)	0.00 (0.14; 0.02)	0.12 (0.17; 0.03)	1.65 (0.87; 0.88)
FS3a	480	0.10 (0.01; 0.00)	-0.01 (0.15; 0.02)	0.22 (0.24; 0.07)	1.97 (0.57; 0.33)
FS6a	384				
	simulation: dom; estimation: rec				
FS1a	640	0.10 (0.01; 0.00)	-0.01 (0.15; 0.02)	0.42 (0.14; 0.12)	-9.71 ($\geq$ 10; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	0.27 (0.37; 0.21)	0.13 (0.15; 0.02)	-0.43 (5.30; $\geq$ 10)
FS6a	384	0.10 (0.01; 0.00)	0.27 (0.38; 0.21)	0.14 (0.16; 0.03)	-0.82 (4.83; $\geq$ 10)
	simulation: dom; estimation: add				
FS1a	640	0.10 (0.01; 0.00)	-0.00 (0.10; 0.01)	0.07 (0.12; 0.01)	2.13 (0.63; 0.42)
FS3a	480	0.10 (0.01; 0.00)	-0.00 (0.10; 0.01)	0.14 (0.17; 0.03)	2.38 (0.53; 0.42)
FS6a	384	0.10 (0.01; 0.00)	-0.01 (0.07; 0.00)	0.10 (0.11; 0.01)	2.42 (0.40; 0.33)
	simulation: rec; estimation: dom				
FS1a	640	0.10 (0.01; 0.00)	-0.01 (0.06; 0.00)	0.14 (0.11; 0.01)	8.50 (5.15; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	-0.02 (0.14; 0.02)	0.59 (0.25; 0.30)	2.78 (3.43; $\geq$ 10)
FS6a	384				
	simulation: rec; estimation: rec				
FS1a	640	0.10 (0.01; 0.00)	0.18 (0.35; 0.16)	0.62 (0.34; 0.38)	-14.31 ($\geq$ 10; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	0.19 (0.38; 0.18)	0.28 (0.28; 0.11)	1.05 (7.10; $\geq$ 10)
FS6a	384	0.10 (0.01; 0.00)	0.18 (0.38; 0.18)	0.28 (0.27; 0.11)	0.24 (6.54; $\geq$ 10)
	simulation: rec; estimation: add				
FS1a	640	0.10 (0.01; 0.00)	-0.00 (0.05; 0.00)	0.15 (0.06; 0.01)	8.46 (5.05; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	-0.01 (0.08; 0.01)	0.41 (0.22; 0.14)	6.75 (6.47; $\geq$ 10)
FS6a	384	0.10 (0.01; 0.00)	-0.01 (0.08; 0.01)	0.42 (0.18; 0.14)	4.97 (5.24; $\geq$ 10)
	simulation: add; estimation: dom				
FS1a	640	0.10 (0.01; 0.00)	-0.01 (0.18; 0.03)	0.19 (0.29; 0.09)	1.50 (0.96; 1.17)
FS3a	480	0.10 (0.01; 0.00)	-0.02 (0.14; 0.02)	0.31 (0.32; 0.15)	1.52 (0.75; 0.79)
FS6a	384				
	simulation: add; estimation: rec				
FS1a	640	0.10 (0.01; 0.00)	-0.02 (0.14; 0.02)	0.42 (0.16; 0.13)	-10.88 (7.59; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	0.26 (0.38; 0.21)	0.16 (0.20; 0.04)	-0.50 (4.18; $\geq$ 10)
FS6a	384	0.10 (0.01; 0.00)	0.27 (0.39; 0.23)	0.17 (0.21; 0.05)	-0.65 (5.09; $\geq$ 10)
	simulation: add; estimation: add				
FS1a	640	0.10 (0.01; 0.00)	-0.01 (0.10; 0.01)	0.10 (0.19; 0.04)	1.96 (0.76; 0.58)
FS6a	384	0.10 (0.01; 0.00)	-0.00 (0.07; 0.01)	0.16 (0.15; 0.03)	1.94 (0.59; 0.35)

Table C.21: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.5$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1a	640	0.10 (0.01; 0.00)	0.71 (0.13; 0.06)	0.09 (0.03; 0.00)	1.38 (0.21; 0.42)
FS6a	384	0.10 (0.01; 0.00)	0.52 (0.10; 0.01)	0.11 (0.04; 0.00)	2.00 (0.23; 0.05)
	simulation: dom; estimation: rec				
FS1a	640	0.10 (0.01; 0.00)	0.10 (0.47; 0.39)	0.32 (0.24; 0.10)	-4.61 ($\geq$ 10; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	0.36 (0.29; 0.10)	0.06 (0.09; 0.01)	1.00 (2.17; 5.70)
FS6a	384	0.10 (0.01; 0.00)	0.37 (0.30; 0.11)	0.07 (0.10; 0.01)	0.94 (1.87; 4.61)
	simulation: dom; estimation: add				
FS1a	640	0.11 (0.01; 0.00)	0.59 (0.10; 0.02)	0.07 (0.02; 0.00)	1.75 (0.19; 0.10)
FS3a	480	0.10 (0.01; 0.00)	0.47 (0.08; 0.01)	0.08 (0.03; 0.00)	2.35 (0.20; 0.16)
FS6a	384	0.10 (0.01; 0.00)	0.47 (0.08; 0.01)	0.08 (0.03; 0.00)	2.34 (0.20; 0.16)
	simulation: rec; estimation: dom				
FS1a	640	0.12 (0.01; 0.00)	0.38 (0.10; 0.02)	0.25 (0.09; 0.03)	3.05 (1.96; 4.94)
FS3a	480	0.11 (0.01; 0.00)	0.41 (0.04; 0.01)	0.41 (0.05; 0.10)	2.33 (0.26; 0.18)
FS6a	384	0.11 (0.01; 0.00)	0.40 (0.04; 0.01)	0.40 (0.05; 0.10)	2.32 (0.25; 0.17)
	simulation: rec; estimation: rec				
FS1a	640	0.10 (0.01; 0.00)	0.67 (0.29; 0.11)	0.12 (0.18; 0.03)	1.65 (8.20; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	0.57 (0.24; 0.06)	0.15 (0.13; 0.02)	2.71 (3.71; $\geq$ 10)
FS6a	384	0.10 (0.01; 0.00)	0.56 (0.24; 0.06)	0.14 (0.12; 0.02)	2.81 (4.18; $\geq$ 10)
	simulation: rec; estimation: add				
FS1a	640	0.12 (0.01; 0.00)	0.45 (0.08; 0.01)	0.23 (0.05; 0.02)	3.32 (1.31; 3.44)
FS3a	480	0.11 (0.01; 0.00)	0.45 (0.04; 0.00)	0.35 (0.04; 0.07)	3.01 (1.28; 2.66)
FS6a	384	0.11 (0.01; 0.00)	0.44 (0.04; 0.00)	0.36 (0.04; 0.07)	2.83 (0.37; 0.83)
	simulation: add; estimation: dom				
FS1a	640	0.10 (0.01; 0.00)	0.66 (0.15; 0.05)	0.08 (0.04; 0.00)	1.35 (0.29; 0.50)
FS3a	480	0.10 (0.01; 0.00)	0.42 (0.08; 0.01)	0.09 (0.04; 0.00)	1.99 (0.31; 0.09)
FS6a	384	0.10 (0.01; 0.00)	0.43 (0.08; 0.01)	0.09 (0.05; 0.00)	1.93 (0.32; 0.11)
	simulation: add; estimation: rec				
FS1a	640	0.10 (0.01; 0.00)	0.38 (0.52; 0.29)	0.27 (0.23; 0.08)	-4.92 ($\geq$ 10; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	0.61 (0.33; 0.12)	0.11 (0.10; 0.01)	0.57 (1.69; 4.89)
FS6a	384	0.10 (0.01; 0.00)	0.57 (0.34; 0.12)	0.12 (0.12; 0.01)	0.45 (1.43; 4.43)
	simulation: add; estimation: add				
FS1a	640	0.10 (0.01; 0.00)	0.72 (0.12; 0.06)	0.10 (0.03; 0.00)	1.41 (0.21; 0.39)
FS3a	480	0.10 (0.01; 0.00)	0.51 (0.08; 0.01)	0.11 (0.04; 0.00)	2.00 (0.26; 0.07)
FS6a	384	0.10 (0.01; 0.00)	0.51 (0.08; 0.01)	0.11 (0.04; 0.00)	1.99 (0.26; 0.07)

Table C.22: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.9$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS3a	480	0.10 (0.01; 0.00)	0.89 (0.07; 0.00)	0.10 (0.02; 0.00)	2.02 (0.13; 0.02)
FS6a	384	0.10 (0.01; 0.00)	0.89 (0.07; 0.01)	0.10 (0.02; 0.00)	2.03 (0.13; 0.02)
	simulation: dom; estimation: rec				
FS1a	640	0.12 (0.01; 0.00)	0.42 (0.56; 0.55)	0.22 (0.24; 0.07)	-9.20 ($\geq$ 10; $\geq$ 10)
FS3a	480	0.11 (0.01; 0.00)	0.41 (0.27; 0.32)	0.05 (0.05; 0.01)	1.04 (0.86; 1.65)
FS6a	384	0.11 (0.01; 0.00)	0.38 (0.25; 0.34)	0.04 (0.04; 0.01)	1.13 (0.76; 1.34)
	simulation: dom; estimation: add				
FS1a	640	0.12 (0.01; 0.00)	0.83 (0.08; 0.01)	0.09 (0.01; 0.00)	1.78 (0.13; 0.06)
FS3a	480	0.11 (0.01; 0.00)	0.71 (0.07; 0.04)	0.07 (0.01; 0.00)	2.42 (0.13; 0.20)
FS6a	384	0.11 (0.01; 0.00)	0.74 (0.07; 0.03)	0.07 (0.01; 0.00)	2.40 (0.14; 0.18)
	simulation: rec; estimation: dom				
FS3a	480	0.13 (0.01; 0.00)	0.51 (0.03; 0.15)	0.36 (0.04; 0.07)	2.98 (2.10; 5.36)
FS6a	384	0.13 (0.01; 0.00)	0.51 (0.03; 0.15)	0.36 (0.04; 0.07)	2.88 (1.85; 4.17)
	simulation: rec; estimation: rec				
FS1a	640	0.10 (0.01; 0.00)	0.91 (0.12; 0.01)	0.09 (0.02; 0.00)	2.08 (1.03; 1.06)
FS3a	480	0.10 (0.01; 0.00)	0.84 (0.14; 0.02)	0.10 (0.03; 0.00)	2.60 (2.12; 4.86)
FS6a	384	0.10 (0.01; 0.00)	0.84 (0.14; 0.02)	0.10 (0.03; 0.00)	2.42 (1.35; 2.01)
	simulation: rec; estimation: add				
FS1a	640	0.14 (0.01; 0.00)	0.60 (0.07; 0.09)	0.26 (0.04; 0.03)	3.99 (3.12; $\geq$ 10)
FS3a	480	0.13 (0.01; 0.00)	0.58 (0.03; 0.10)	0.29 (0.03; 0.04)	4.61 (4.11; $\geq$ 10)
FS6a	384	0.13 (0.01; 0.00)	0.58 (0.03; 0.10)	0.29 (0.03; 0.04)	4.21 (3.60; $\geq$ 10)
	simulation: add; estimation: dom				
FS1a	640	0.11 (0.01; 0.00)	0.93 (0.07; 0.01)	0.09 (0.01; 0.00)	1.45 (0.13; 0.33)
FS3a	480	0.11 (0.01; 0.00)	0.69 (0.07; 0.05)	0.06 (0.01; 0.00)	2.20 (0.15; 0.06)
FS6a	384	0.11 (0.01; 0.00)	0.71 (0.07; 0.04)	0.07 (0.01; 0.00)	2.15 (0.16; 0.05)
	simulation: add; estimation: rec				
FS1a	640	0.11 (0.01; 0.00)	0.94 (0.19; 0.04)	0.11 (0.07; 0.00)	-1.10 ($\geq$ 10; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	0.80 (0.25; 0.07)	0.10 (0.04; 0.00)	0.52 (0.80; 2.81)
FS6a	384	0.10 (0.01; 0.00)	0.81 (0.25; 0.07)	0.10 (0.05; 0.00)	0.52 (0.40; 2.35)
	simulation: add; estimation: add				
FS1a	640	0.10 (0.01; 0.00)	0.98 (0.02; 0.01)	0.10 (0.01; 0.00)	1.57 (0.07; 0.19)
FS3a	480	0.10 (0.01; 0.00)	0.89 (0.06; 0.00)	0.10 (0.01; 0.00)	2.02 (0.13; 0.02)
FS6a	384	0.10 (0.01; 0.00)	0.89 (0.06; 0.00)	0.10 (0.01; 0.00)	2.02 (0.12; 0.02)

Table C.23: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.0$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1a	640	0.20 (0.01; 0.00)	-0.00 (0.13; 0.02)	0.15 (0.16; 0.03)	1.74 (0.52; 0.34)
FS3a	480	0.20 (0.01; 0.00)	-0.00 (0.17; 0.03)	0.30 (0.25; 0.07)	2.00 (0.41; 0.17)
FS6a	384	0.20 (0.01; 0.00)	-0.01 (0.17; 0.03)	0.29 (0.23; 0.06)	1.98 (0.37; 0.14)
	simulation: dom; estimation: rec				
FS1a	640	0.20 (0.01; 0.00)	0.03 (0.29; 0.09)	0.30 (0.16; 0.03)	-5.78 (6.94; $\geq$ 10)
FS3a	480	0.20 (0.01; 0.00)	-0.00 (0.32; 0.10)	0.21 (0.21; 0.05)	1.32 (1.67; 3.24)
FS6a	384	0.20 (0.01; 0.00)	0.01 (0.31; 0.10)	0.20 (0.21; 0.05)	1.15 (1.44; 2.79)
	simulation: dom; estimation: add				
FS1a	640	0.20 (0.01; 0.00)	0.00 (0.04; 0.00)	0.07 (0.05; 0.02)	2.45 (0.36; 0.33)
FS3a	480	0.20 (0.01; 0.00)	-0.00 (0.08; 0.01)	0.18 (0.13; 0.02)	2.61 (0.34; 0.48)
FS6a	384	0.20 (0.01; 0.00)	-0.00 (0.05; 0.00)	0.14 (0.07; 0.01)	2.68 (0.27; 0.53)
	simulation: rec; estimation: dom				
FS1a	640	0.20 (0.01; 0.00)	-0.00 (0.07; 0.01)	0.26 (0.18; 0.04)	3.80 (2.63; $\geq$ 10)
FS3a	480	0.20 (0.01; 0.00)	-0.02 (0.13; 0.02)	0.73 (0.18; 0.31)	1.86 (0.40; 0.18)
FS6a	384	0.20 (0.01; 0.00)	-0.02 (0.12; 0.01)	0.72 (0.16; 0.30)	1.85 (0.33; 0.13)
	simulation: rec; estimation: rec				
FS1a	640	0.20 (0.01; 0.00)	0.10 (0.27; 0.08)	0.51 (0.23; 0.15)	-9.76 ($\geq$ 10; $\geq$ 10)
FS3a	480	0.20 (0.01; 0.00)	-0.00 (0.24; 0.06)	0.37 (0.32; 0.13)	2.87 (3.51; $\geq$ 10)
FS6a	384	0.20 (0.01; 0.00)	-0.00 (0.26; 0.07)	0.34 (0.28; 0.10)	2.59 (3.50; $\geq$ 10)
	simulation: rec; estimation: add				
FS1a	640	0.20 (0.01; 0.00)	0.00 (0.05; 0.00)	0.25 (0.08; 0.01)	4.04 (1.80; 7.39)
FS3a	480	0.20 (0.01; 0.00)	0.00 (0.06; 0.00)	0.59 (0.12; 0.17)	2.59 (0.36; 0.47)
FS6a	384	0.20 (0.01; 0.00)	0.00 (0.06; 0.00)	0.57 (0.09; 0.14)	2.61 (0.28; 0.45)
	simulation: add; estimation: dom				
FS1a	640	0.20 (0.01; 0.00)	-0.01 (0.13; 0.02)	0.26 (0.30; 0.09)	1.36 (0.67; 0.85)
FS3a	480	0.20 (0.01; 0.00)	-0.02 (0.14; 0.02)	0.48 (0.32; 0.18)	1.30 (0.59; 0.83)
FS6a	384	0.20 (0.01; 0.00)	-0.01 (0.13; 0.02)	0.47 (0.27; 0.14)	1.24 (0.46; 0.79)
	simulation: add; estimation: rec				
FS1a	640	0.20 (0.01; 0.00)	0.03 (0.29; 0.08)	0.34 (0.20; 0.06)	-5.79 (5.82; $\geq$ 10)
FS3a	480	0.20 (0.01; 0.00)	0.01 (0.31; 0.09)	0.25 (0.28; 0.08)	1.17 (1.69; 3.53)
FS6a	384	0.20 (0.01; 0.00)	0.01 (0.29; 0.09)	0.22 (0.27; 0.07)	1.23 (1.81; 3.86)
	simulation: add; estimation: add				
FS1a	640	0.20 (0.01; 0.00)	-0.00 (0.05; 0.00)	0.10 (0.11; 0.02)	2.19 (0.58; 0.37)
FS3a	480	0.20 (0.01; 0.00)	0.00 (0.05; 0.00)	0.24 (0.16; 0.03)	2.03 (0.52; 0.27)
FS6a	384	0.20 (0.01; 0.00)	0.00 (0.05; 0.00)	0.23 (0.12; 0.02)	2.03 (0.46; 0.21)

Table C.24: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.5$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1a	640	0.20 (0.01; 0.00)	0.61 (0.12; 0.03)	0.17 (0.06; 0.00)	1.44 (0.19; 0.34)
FS3a	480	0.20 (0.01; 0.00)	0.51 (0.11; 0.01)	0.21 (0.07; 0.01)	2.01 (0.21; 0.04)
FS6a	384	0.20 (0.01; 0.00)	0.51 (0.11; 0.01)	0.21 (0.07; 0.01)	2.00 (0.20; 0.04)
	simulation: dom; estimation: rec				
FS1a	640	0.22 (0.01; 0.00)	0.40 (0.42; 0.19)	0.14 (0.17; 0.03)	0.33 (1.42; 4.80)
FS3a	480	0.21 (0.01; 0.00)	0.31 (0.12; 0.05)	0.04 (0.03; 0.03)	1.44 (0.48; 0.54)
FS6a	384	0.21 (0.01; 0.00)	0.31 (0.13; 0.05)	0.05 (0.05; 0.03)	1.42 (0.50; 0.59)
	simulation: dom; estimation: add				
FS1a	640	0.21 (0.01; 0.00)	0.46 (0.07; 0.01)	0.11 (0.03; 0.01)	2.08 (0.15; 0.03)
FS3a	480	0.21 (0.01; 0.00)	0.43 (0.06; 0.01)	0.14 (0.03; 0.00)	2.63 (0.16; 0.43)
FS6a	384	0.20 (0.01; 0.00)	0.43 (0.06; 0.01)	0.14 (0.03; 0.00)	2.62 (0.17; 0.42)
	simulation: rec; estimation: dom				
FS1a	640	0.22 (0.01; 0.00)	0.34 (0.11; 0.04)	0.34 (0.13; 0.04)	2.53 (0.81; 0.93)
FS3a	480	0.21 (0.01; 0.00)	0.40 (0.05; 0.01)	0.56 (0.07; 0.13)	2.19 (0.23; 0.09)
FS6a	384	0.21 (0.01; 0.00)	0.40 (0.05; 0.01)	0.57 (0.06; 0.14)	2.17 (0.20; 0.07)
	simulation: rec; estimation: rec				
FS1a	640	0.21 (0.01; 0.00)	0.64 (0.21; 0.06)	0.15 (0.08; 0.01)	2.59 (2.63; 7.24)
FS3a	480	0.20 (0.01; 0.00)	0.53 (0.14; 0.02)	0.23 (0.10; 0.01)	2.12 (1.23; 1.53)
FS6a	384	0.20 (0.01; 0.00)	0.54 (0.15; 0.02)	0.23 (0.10; 0.01)	2.07 (1.28; 1.64)
	simulation: rec; estimation: add				
FS1a	640	0.21 (0.01; 0.00)	0.43 (0.07; 0.01)	0.31 (0.06; 0.01)	3.20 (0.55; 1.74)
FS3a	480	0.21 (0.01; 0.00)	0.46 (0.04; 0.00)	0.48 (0.05; 0.08)	2.97 (0.24; 1.00)
FS6a	384	0.21 (0.01; 0.00)	0.46 (0.04; 0.00)	0.48 (0.05; 0.08)	2.92 (0.22; 0.90)
	simulation: add; estimation: dom				
FS1a	640	0.20 (0.01; 0.00)	0.53 (0.15; 0.02)	0.16 (0.09; 0.01)	1.30 (0.34; 0.60)
FS3a	480	0.20 (0.01; 0.00)	0.37 (0.09; 0.02)	0.17 (0.11; 0.01)	1.89 (0.41; 0.18)
FS6a	384	0.20 (0.01; 0.00)	0.41 (0.10; 0.02)	0.21 (0.12; 0.02)	1.74 (0.42; 0.25)
	simulation: add; estimation: rec				
FS1a	640	0.21 (0.01; 0.00)	0.86 (0.20; 0.17)	0.18 (0.06; 0.00)	0.20 (1.00; 4.24)
FS3a	480	0.20 (0.01; 0.00)	0.69 (0.25; 0.10)	0.18 (0.10; 0.01)	0.70 (0.45; 1.89)
FS6a	384	0.20 (0.01; 0.00)	0.64 (0.26; 0.09)	0.16 (0.09; 0.01)	0.79 (0.47; 1.68)
	simulation: add; estimation: add				
FS1a	640	0.19 (0.01; 0.00)	0.64 (0.11; 0.03)	0.19 (0.05; 0.00)	1.48 (0.20; 0.32)
FS6a	384	0.20 (0.01; 0.00)	0.51 (0.07; 0.00)	0.21 (0.05; 0.00)	2.00 (0.23; 0.05)

Table C.25: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.9$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1a	640	0.21 (0.01; 0.00)	0.98 (0.03; 0.01)	0.20 (0.01; 0.00)	1.48 (0.07; 0.27)
FS6a	384	0.20 (0.01; 0.00)	0.89 (0.07; 0.00)	0.20 (0.03; 0.00)	2.01 (0.11; 0.01)
	simulation: dom; estimation: rec				
FS1a	640	0.25 (0.01; 0.00)	0.38 (0.19; 0.31)	0.06 (0.10; 0.03)	0.95 (0.50; 1.36)
FS3a	480	0.23 (0.01; 0.00)	0.31 (0.08; 0.36)	0.03 (0.02; 0.03)	1.61 (0.32; 0.25)
FS6a	384	0.23 (0.01; 0.00)	0.31 (0.09; 0.36)	0.03 (0.02; 0.03)	1.64 (0.35; 0.25)
	simulation: dom; estimation: add				
FS1a	640	0.23 (0.01; 0.00)	0.66 (0.06; 0.06)	0.12 (0.02; 0.01)	2.11 (0.12; 0.02)
FS3a	480	0.22 (0.01; 0.00)	0.62 (0.05; 0.08)	0.11 (0.01; 0.01)	2.76 (0.12; 0.58)
FS6a	384	0.22 (0.01; 0.00)	0.64 (0.05; 0.07)	0.12 (0.02; 0.01)	2.72 (0.12; 0.54)
	simulation: rec; estimation: dom				
FS1a	640	0.25 (0.01; 0.00)	0.50 (0.10; 0.17)	0.46 (0.11; 0.08)	2.06 (0.61; 0.37)
FS6a	384	0.23 (0.01; 0.00)	0.53 (0.03; 0.14)	0.51 (0.03; 0.10)	2.39 (0.13; 0.17)
	simulation: rec; estimation: rec				
FS1a	640	0.21 (0.01; 0.00)	0.94 (0.08; 0.01)	0.19 (0.03; 0.00)	1.81 (0.34; 0.15)
FS3a	480	0.20 (0.01; 0.00)	0.88 (0.09; 0.01)	0.19 (0.03; 0.00)	2.13 (0.44; 0.21)
FS6a	384	0.20 (0.01; 0.00)	0.88 (0.09; 0.01)	0.20 (0.03; 0.00)	2.10 (0.38; 0.15)
	simulation: rec; estimation: add				
FS1a	640	0.22 (0.01; 0.00)	0.67 (0.05; 0.06)	0.35 (0.03; 0.02)	3.01 (0.31; 1.11)
FS3a	480	0.22 (0.01; 0.00)	0.65 (0.02; 0.06)	0.40 (0.02; 0.04)	3.47 (0.16; 2.17)
FS6a	384	0.22 (0.01; 0.00)	0.64 (0.02; 0.07)	0.41 (0.02; 0.04)	3.42 (0.15; 2.05)
	simulation: add; estimation: dom				
FS1a	640	0.20 (0.01; 0.00)	0.90 (0.15; 0.02)	0.17 (0.04; 0.00)	1.31 (0.25; 0.53)
FS6a	384	0.21 (0.01; 0.00)	0.59 (0.05; 0.10)	0.10 (0.02; 0.01)	2.23 (0.15; 0.08)
	simulation: add; estimation: rec				
FS1a	640	0.22 (0.01; 0.00)	0.96 (0.02; 0.00)	0.21 (0.01; 0.00)	0.45 (0.07; 2.41)
FS3a	480	0.20 (0.01; 0.00)	0.96 (0.06; 0.01)	0.20 (0.02; 0.00)	0.63 (0.09; 1.90)
FS6a	384	0.20 (0.01; 0.00)	0.94 (0.09; 0.01)	0.20 (0.03; 0.00)	0.64 (0.12; 1.86)
	simulation: add; estimation: add				
FS1a	640	0.19 (0.01; 0.00)	0.99 (0.01; 0.01)	0.19 (0.01; 0.00)	1.55 (0.06; 0.21)
FS3a	480	0.20 (0.01; 0.00)	0.89 (0.05; 0.00)	0.20 (0.02; 0.00)	2.02 (0.11; 0.01)
FS6a	384	0.20 (0.01; 0.00)	0.90 (0.05; 0.00)	0.20 (0.02; 0.00)	2.01 (0.11; 0.01)

Table C.26: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.1$	$R_1 = 0.5$	$p^\star = 0.2$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS1a	640	0.10 (0.01; 0.00)	0.58 (0.10; 0.02)	0.16 (0.05; 0.00)	1.48 (0.19; 0.31)
FS3a	480	0.10 (0.01; 0.00)	0.49 (0.09; 0.01)	0.19 (0.06; 0.00)	2.04 (0.20; 0.04)
FS6a	384	0.10 (0.01; 0.00)	0.49 (0.09; 0.01)	0.20 (0.06; 0.00)	2.03 (0.19; 0.04)
	simulation: dom; estimation: rec				
FS1a	640	0.11 (0.01; 0.00)	0.62 (0.43; 0.20)	0.12 (0.15; 0.03)	0.22 (1.48; 5.34)
FS3a	480	0.11 (0.01; 0.00)	0.35 (0.17; 0.05)	0.05 (0.04; 0.03)	1.42 (0.78; 0.94)
FS6a	384	0.10 (0.01; 0.00)	0.37 (0.19; 0.05)	0.05 (0.05; 0.02)	1.40 (0.75; 0.91)
	simulation: dom; estimation: add				
FS1a	640	0.10 (0.01; 0.00)	0.48 (0.08; 0.01)	0.11 (0.03; 0.01)	2.07 (0.16; 0.03)
FS3a	480	0.10 (0.01; 0.00)	0.46 (0.07; 0.01)	0.15 (0.04; 0.00)	2.59 (0.17; 0.38)
FS6a	384	0.10 (0.01; 0.00)	0.46 (0.07; 0.01)	0.14 (0.03; 0.00)	2.60 (0.16; 0.38)
	simulation: rec; estimation: dom				
FS1a	640	0.11 (0.01; 0.00)	0.32 (0.08; 0.04)	0.30 (0.09; 0.02)	2.83 (1.27; 2.31)
FS3a	480	0.11 (0.01; 0.00)	0.34 (0.03; 0.03)	0.50 (0.05; 0.09)	2.36 (0.21; 0.17)
FS6a	384	0.11 (0.01; 0.00)	0.33 (0.04; 0.03)	0.51 (0.06; 0.10)	2.33 (0.20; 0.15)
	simulation: rec; estimation: rec				
FS1a	640	0.11 (0.01; 0.00)	0.60 (0.18; 0.04)	0.12 (0.06; 0.01)	3.37 (4.73; $\geq$ 10)
FS3a	480	0.10 (0.01; 0.00)	0.50 (0.12; 0.02)	0.21 (0.08; 0.01)	2.29 (1.49; 2.31)
FS6a	384	0.10 (0.01; 0.00)	0.50 (0.13; 0.02)	0.20 (0.08; 0.01)	2.20 (1.06; 1.17)
	simulation: rec; estimation: add				
FS1a	640	0.10 (0.01; 0.00)	0.43 (0.06; 0.01)	0.28 (0.04; 0.01)	3.47 (0.60; 2.50)
FS3a	480	0.11 (0.01; 0.00)	0.39 (0.03; 0.01)	0.43 (0.04; 0.05)	3.21 (0.22; 1.52)
FS6a	384	0.11 (0.01; 0.00)	0.39 (0.03; 0.01)	0.44 (0.04; 0.06)	3.15 (0.20; 1.36)
	simulation: add; estimation: dom				
FS1a	640	0.10 (0.01; 0.00)	0.51 (0.11; 0.01)	0.16 (0.07; 0.01)	1.29 (0.31; 0.60)
FS3a	480	0.10 (0.01; 0.00)	0.37 (0.08; 0.02)	0.17 (0.09; 0.01)	1.86 (0.39; 0.17)
FS6a	384	0.10 (0.01; 0.00)	0.39 (0.08; 0.02)	0.21 (0.10; 0.01)	1.71 (0.39; 0.23)
	simulation: add; estimation: rec				
FS1a	640	0.10 (0.01; 0.00)	0.88 (0.27; 0.22)	0.12 (0.09; 0.01)	0.24 (1.53; 5.44)
FS3a	480	0.10 (0.01; 0.00)	0.62 (0.21; 0.06)	0.15 (0.08; 0.01)	0.77 (0.48; 1.74)
FS6a	384	0.10 (0.01; 0.00)	0.59 (0.22; 0.06)	0.14 (0.08; 0.01)	0.83 (0.63; 1.76)
	simulation: add; estimation: add				
FS3a	480	0.10 (0.01; 0.00)	0.50 (0.06; 0.00)	0.20 (0.05; 0.00)	2.02 (0.23; 0.05)
FS6a	384	0.10 (0.01; 0.00)	0.51 (0.06; 0.00)	0.20 (0.05; 0.00)	2.02 (0.22; 0.05)

Table C.27: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$R_1 = 0.5$	$p^\star = 0.1$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS3a	480	0.20 (0.01; 0.00)	0.52 (0.10; 0.01)	0.11 (0.04; 0.00)	1.99 (0.21; 0.05)
FS6a	384	0.20 (0.01; 0.00)	0.51 (0.10; 0.01)	0.11 (0.04; 0.00)	2.00 (0.22; 0.05)
	simulation: dom; estimation: rec				
FS3a	480	0.21 (0.01; 0.00)	0.32 (0.25; 0.09)	0.07 (0.10; 0.01)	0.80 (2.84; 9.52)
FS6a	384	0.20 (0.01; 0.00)	0.32 (0.25; 0.09)	0.08 (0.11; 0.01)	0.82 (1.98; 5.29)
	simulation: dom; estimation: add				
FS3a	480	0.21 (0.01; 0.00)	0.44 (0.07; 0.01)	0.08 (0.02; 0.00)	2.38 (0.20; 0.18)
FS6a	384	0.21 (0.01; 0.00)	0.45 (0.07; 0.01)	0.08 (0.02; 0.00)	2.36 (0.21; 0.17)
	simulation: rec; estimation: dom				
FS3a	480	0.21 (0.01; 0.00)	0.47 (0.06; 0.01)	0.47 (0.07; 0.14)	2.11 (0.30; 0.10)
FS6a	384	0.21 (0.01; 0.00)	0.46 (0.06; 0.01)	0.46 (0.07; 0.13)	2.14 (0.29; 0.10)
	simulation: rec; estimation: rec				
FS3a	480	0.20 (0.01; 0.00)	0.56 (0.22; 0.05)	0.16 (0.13; 0.02)	2.58 (3.15; $\geq$ 10)
FS6a	384	0.20 (0.01; 0.00)	0.56 (0.22; 0.05)	0.16 (0.13; 0.02)	2.36 (2.81; 8.00)
	simulation: rec; estimation: add				
FS3a	480	0.21 (0.01; 0.00)	0.50 (0.06; 0.00)	0.40 (0.06; 0.09)	2.66 (0.66; 0.87)
FS6a	384	0.21 (0.01; 0.00)	0.50 (0.06; 0.00)	0.40 (0.06; 0.09)	2.61 (0.31; 0.46)
	simulation: add; estimation: dom				
FS3a	480	0.20 (0.01; 0.00)	0.42 (0.07; 0.01)	0.08 (0.04; 0.00)	2.03 (0.27; 0.07)
FS6a	384	0.20 (0.01; 0.00)	0.44 (0.09; 0.01)	0.09 (0.05; 0.00)	1.97 (0.31; 0.10)
	simulation: add; estimation: rec				
FS3a	480	0.20 (0.01; 0.00)	0.53 (0.26; 0.07)	0.13 (0.11; 0.01)	0.51 (1.38; 4.13)
FS6a	384	0.20 (0.01; 0.00)	0.50 (0.27; 0.08)	0.14 (0.13; 0.02)	0.48 (1.20; 3.75)
	simulation: add; estimation: add				
FS3a	480	0.20 (0.01; 0.00)	0.51 (0.08; 0.01)	0.11 (0.04; 0.00)	1.99 (0.25; 0.06)
FS6a	384	0.20 (0.01; 0.00)	0.51 (0.08; 0.01)	0.10 (0.03; 0.00)	2.01 (0.24; 0.06)

Table C.28: Parameter estimates for a total of N = 1280 individuals and ascertainment on one affected offspring. Standard deviation and MSE are given in parentheses. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.3$	$R_1 = 0.9$	$p^\star = 0.3$	$\beta = 2.0$
	simulation: dom; estimation: dom				
FS3a	480	0.30 (0.01; 0.00)	0.88 (0.07; 0.01)	0.29 (0.04; 0.00)	2.02 (0.11; 0.01)
FS6a	384	0.30 (0.01; 0.00)	0.88 (0.07; 0.01)	0.29 (0.04; 0.00)	2.02 (0.10; 0.01)
	simulation: dom; estimation: rec				
FS3a	480				
FS6a	384	0.34 (0.01; 0.00)	0.32 (0.05; 0.34)	0.05 (0.01; 0.06)	1.82 (0.23; 0.08)
	simulation: dom; estimation: add				
FS3a	480				
FS6a	384	0.32 (0.01; 0.00)	0.58 (0.04; 0.10)	0.15 (0.02; 0.02)	3.01 (0.11; 1.04)
	simulation: rec; estimation: dom				
FS3a	480	0.32 (0.01; 0.00)	0.53 (0.03; 0.14)	0.62 (0.04; 0.10)	2.27 (0.14; 0.09)
FS6a	384	0.32 (0.01; 0.00)	0.52 (0.04; 0.14)	0.62 (0.04; 0.10)	2.26 (0.13; 0.09)
	simulation: rec; estimation: rec				
FS3a	480	0.30 (0.01; 0.00)	0.88 (0.08; 0.01)	0.29 (0.04; 0.00)	2.09 (0.30; 0.10)
FS6a	384	0.30 (0.01; 0.00)	0.88 (0.08; 0.01)	0.29 (0.04; 0.00)	2.08 (0.30; 0.09)
	simulation: rec; estimation: add				
FS3a	480				
FS6a	384	0.31 (0.01; 0.00)	0.68 (0.02; 0.05)	0.49 (0.02; 0.04)	3.47 (0.13; 2.17)
	simulation: add; estimation: dom				
FS3a	480				
FS6a	384	0.30 (0.01; 0.00)	0.50 (0.06; 0.17)	0.13 (0.05; 0.03)	2.24 (0.23; 0.11)
	simulation: add; estimation: rec				
FS3a	480	0.31 (0.01; 0.00)	0.98 (0.02; 0.01)	0.30 (0.01; 0.00)	0.78 (0.06; 1.50)
FS6a	384	0.31 (0.01; 0.00)	0.97 (0.03; 0.01)	0.30 (0.02; 0.00)	0.78 (0.07; 1.50)
	simulation: add; estimation: add				
FS3a	480	0.30 (0.01; 0.00)	0.90 (0.05; 0.00)	0.30 (0.03; 0.00)	2.01 (0.10; 0.01)
FS6a	384	0.30 (0.01; 0.00)	0.90 (0.05; 0.00)	0.30 (0.03; 0.00)	2.01 (0.10; 0.01)

C.4. **Simulations for the two locus case.** All tables contain results for misspecifiation, *i.e.* simulation and estimation iterates all possible combinations for recessive, additive and dominant penetrance models.

Table C.29: Summary of parameter combinations simulated in two locus case. Column A specifies whether samples were ascertained. n is the total number of individuals.

Table	Page	A	n	η_1	η_2	η_3	R_1	R_2	$R_3 1$	$p^\star$	β
C.30	168	-	1800	0.3	0.1	0.3	0.3	0.1	-0.1	0.3	2.0
C.31	171	-	1800	0.2	0.3	0.1	0.3	0.1	-0.1	0.3	2.0
C.32	173	-	1800	0.2	0.2	0.3	0.1	0.2	0.0	0.3	2.0
C.33	175	+	1920	0.3	0.1	0.3	0.3	0.1	-0.1	0.3	2.0
C.34	177	+	1920	0.2	0.3	0.1	0.3	0.1	-0.1	0.3	2.0
C.35	179	+	1920	0.2	0.2	0.3	0.1	0.2	0.0	0.3	2.0

Table C.30: Parameter estimates for a total of N = 1800 individuals and unconditional sampling. MSE is given in parentheses in a separate line. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.3$	$\eta_2 = 0.1$	$\eta_3 = 0.3$	$R_1 = 0.3$	$R_2 = 0.1$	$R_3 = -0.1$	$p^\star = 0.3$	$\beta = 2.0$
	simulation: dom; estimation: dom								
FS1	900	0.29	0.12	0.26	0.40	0.13	-0.19	0.32	1.32
		(0.00)	(0.00)	(0.00)	(0.03)	(0.02)	(0.02)	(0.00)	(0.47)
FS3	600	0.29	0.12	0.26	0.32	0.10	-0.15	0.32	1.97
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.05)
FS4	600	0.29	0.12	0.26	0.37	0.12	-0.17	0.32	1.52
		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.01)	(0.00)	(0.25)
FS6	450	0.29	0.12	0.26	0.31	0.10	-0.14	0.32	1.98
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)
	simulation: dom; estimation: rec								
FS1	900	0.29	0.12	0.26	0.28	0.11	-0.08	0.34	-0.02
		(0.00)	(0.00)	(0.00)	(0.09)	(0.08)	(0.07)	(0.01)	(4.15)
FS3	600	0.29	0.12	0.26	0.44	0.15	-0.20	0.32	0.66
		(0.00)	(0.00)	(0.00)	(0.05)	(0.06)	(0.03)	(0.00)	(1.82)
FS4	600	0.29	0.12	0.26	0.42	0.18	-0.17	0.32	0.23
		(0.00)	(0.00)	(0.00)	(0.06)	(0.07)	(0.03)	(0.00)	(3.19)
FS6	450	0.29	0.12	0.26	0.43	0.15	-0.18	0.31	0.66
		(0.00)	(0.00)	(0.00)	(0.05)	(0.05)	(0.03)	(0.00)	(1.82)
	simulation: dom; estimation: add								
FS1	900	0.28	0.12	0.26	0.38	0.12	-0.18	0.25	1.92
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.06)
FS3	600	0.29	0.12	0.26	0.31	0.10	-0.14	0.25	2.65
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.48)
FS4	600	0.29	0.12	0.26	0.33	0.10	-0.16	0.23	2.19
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.08)

FS6	450	0.29	0.12	0.26	0.30	0.10	-0.14	0.24	2.66
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.48)
	simulation: rec; estimation: dom								
FS1	900	0.29	0.12	0.26	0.25	0.09	-0.11	0.53	2.06
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.07)	(0.32)
FS3	600	0.29	0.12	0.26	0.22	0.07	-0.10	0.59	2.66
		(0.00)	(0.00)	(0.00)	(0.02)	(0.01)	(0.01)	(0.10)	(1.14)
FS4	600	0.29	0.12	0.26	0.27	0.09	-0.12	0.60	1.93
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.11)	(0.18)
FS6	450	0.29	0.12	0.26	0.24	0.07	-0.12	0.64	2.32
		(0.00)	(0.00)	(0.00)	(0.02)	(0.01)	(0.01)	(0.13)	(0.42)
	simulation: rec; estimation: rec								
FS1	900	0.29	0.12	0.26	0.47	0.18	-0.22	0.32	0.50
		(0.00)	(0.00)	(0.00)	(0.09)	(0.06)	(0.05)	(0.01)	(3.23)
FS3	600	0.29	0.12	0.26	0.33	0.10	-0.15	0.33	1.96
		(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.02)	(0.01)	(0.58)
FS4	600	0.29	0.12	0.26	0.48	0.16	-0.22	0.31	1.08
		(0.00)	(0.00)	(0.00)	(0.07)	(0.05)	(0.04)	(0.00)	(1.18)
FS6	450	0.29	0.12	0.26	0.33	0.11	-0.16	0.34	1.94
		(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.02)	(0.02)	(0.68)
	simulation: rec; estimation: add								
FS1	900	0.28	0.12	0.26	0.31	0.11	-0.15	0.49	2.72
		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.05)	(0.96)
FS3	600	0.29	0.12	0.26	0.26	0.08	-0.12	0.51	4.06
		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.05)	(9.32)
FS4	600	0.28	0.12	0.26	0.30	0.10	-0.14	0.52	2.71
		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.06)	(0.70)
FS6	450	0.29	0.12	0.26	0.28	0.09	-0.13	0.57	3.13
		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.08)	(1.53)
	simulation: add; estimation: dom								
FS1	900	0.29	0.12	0.26	0.30	0.11	-0.14	0.33	1.03
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.95)
FS3	600	0.29	0.12	0.26	0.25	0.07	-0.11	0.37	1.44
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.36)
FS4	600	0.29	0.12	0.26	0.29	0.10	-0.13	0.35	1.13
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.77)
FS6	450	0.29	0.12	0.26	0.27	0.08	-0.12	0.40	1.36
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.02)	(0.44)
	simulation: add; estimation: rec								
FS1	900	0.29	0.12	0.26	0.28	0.12	-0.11	0.34	0.00
		(0.00)	(0.00)	(0.00)	(0.13)	(0.08)	(0.07)	(0.00)	(4.05)
FS3	600	0.29	0.12	0.26	0.46	0.15	-0.20	0.32	0.53
		(0.00)	(0.00)	(0.00)	(0.05)	(0.05)	(0.03)	(0.00)	(2.17)
FS4	600	0.29	0.12	0.26	0.48	0.16	-0.20	0.32	0.21
		(0.00)	(0.00)	(0.00)	(0.09)	(0.07)	(0.04)	(0.00)	(3.22)
FS6	450	0.29	0.12	0.26	0.45	0.15	-0.19	0.32	0.54

		(0.00)	(0.00)	(0.00)	(0.05)	(0.05)	(0.03)	(0.00)	(2.15)
	simulation: add; estimation: add								
FS1	900	0.28	0.12	0.26	0.39	0.12	-0.18	0.32	1.34
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.45)
FS3	600	0.29	0.12	0.26	0.31	0.10	-0.14	0.32	1.97
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.09)
FS4	600	0.29	0.12	0.26	0.34	0.11	-0.16	0.31	1.52
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.26)
FS6	450	0.29	0.12	0.26	0.31	0.10	-0.14	0.31	1.98
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.07)

Table C.31: Parameter estimates for a total of N = 1800 individuals and unconditional sampling. MSE is given in parentheses in a separate line. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$\eta_2 = 0.2$	$\eta_3 = 0.3$	$R_1 = 0.1$	$R_2 = 0.3$	$R_3 = -0.1$	$p^\star = 0.3$	$\beta = 2.0$
	simulation: dom; estimation: dom								
FS1	900	0.24	0.17	0.26	0.19	0.37	-0.19	0.31	1.36
		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.01)	(0.00)	(0.43)
FS3	600	0.24	0.17	0.26	0.15	0.29	-0.14	0.30	2.01
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.03)
FS4	600	0.24	0.17	0.26	0.17	0.35	-0.17	0.31	1.54
		(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.01)	(0.00)	(0.23)
FS6	450	0.24	0.17	0.26	0.15	0.30	-0.14	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.03)
	simulation: dom; estimation: rec								
FS1	900	0.24	0.17	0.26	0.14	0.26	-0.07	0.32	-0.02
		(0.00)	(0.00)	(0.00)	(0.08)	(0.13)	(0.08)	(0.00)	(4.17)
FS3	600	0.24	0.17	0.26	0.20	0.40	-0.18	0.30	0.68
		(0.00)	(0.00)	(0.00)	(0.06)	(0.07)	(0.03)	(0.00)	(1.76)
FS4	600	0.24	0.17	0.26	0.25	0.41	-0.17	0.31	0.24
		(0.00)	(0.00)	(0.00)	(0.08)	(0.08)	(0.04)	(0.00)	(3.14)
FS6	450	0.24	0.17	0.26	0.21	0.39	-0.18	0.30	0.69
		(0.00)	(0.00)	(0.00)	(0.06)	(0.06)	(0.04)	(0.00)	(1.75)
	simulation: dom; estimation: add								
FS1	900	0.24	0.17	0.26	0.18	0.38	-0.18	0.26	1.85
		(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.01)	(0.00)	(0.06)
FS3	600	0.24	0.17	0.26	0.15	0.31	-0.15	0.26	2.60
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.40)
FS4	600	0.24	0.17	0.26	0.16	0.33	-0.16	0.25	2.10
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.05)
FS6	450	0.24	0.17	0.26	0.14	0.30	-0.14	0.25	2.63
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.43)
	simulation: rec; estimation: dom								
FS1	900	0.24	0.17	0.26	0.11	0.19	-0.11	0.43	2.59
		(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.01)	(0.03)	(0.74)
FS3	600	0.24	0.17	0.26	0.09	0.18	-0.09	0.52	3.21
		(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.01)	(0.07)	(3.30)
FS4	600	0.24	0.17	0.26	0.12	0.22	-0.12	0.54	2.18
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.08)	(0.32)
FS6	450	0.24	0.17	0.26	0.11	0.21	-0.11	0.61	2.47
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.11)	(0.62)
	simulation: rec; estimation: rec								
FS1	900	0.24	0.17	0.26	0.24	0.48	-0.24	0.31	0.51
		(0.00)	(0.00)	(0.00)	(0.07)	(0.11)	(0.05)	(0.00)	(3.80)
FS3	600	0.24	0.17	0.26	0.14	0.30	-0.14	0.31	2.00

		(0.00)	(0.00)	(0.00)	(0.02)	(0.03)	(0.02)	(0.00)	(0.23)
FS4	600	0.24	0.17	0.26	0.23	0.44	-0.22	0.30	1.06
		(0.00)	(0.00)	(0.00)	(0.06)	(0.07)	(0.04)	(0.00)	(1.25)
FS6	450	0.24	0.17	0.26	0.15	0.28	-0.14	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.02)	(0.00)	(0.31)
	simulation: rec; estimation: add								
FS1	900	0.24	0.16	0.26	0.13	0.26	-0.13	0.41	3.39
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.02)	(2.59)
FS3	600	0.24	0.17	0.26	0.12	0.23	-0.12	0.47	4.90
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.04)	($\geq$ 10)
FS4	600	0.24	0.17	0.26	0.14	0.27	-0.14	0.49	2.89
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.04)	(1.08)
FS6	450	0.24	0.17	0.26	0.13	0.25	-0.12	0.54	3.40
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.06)	(2.89)
	simulation: add; estimation: dom								
FS1	900	0.24	0.17	0.26	0.17	0.28	-0.15	0.32	1.06
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.89)
FS3	600	0.24	0.17	0.26	0.11	0.22	-0.11	0.33	1.52
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.25)
FS4	600	0.24	0.17	0.26	0.14	0.26	-0.13	0.33	1.17
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.70)
FS6	450	0.24	0.17	0.26	0.12	0.23	-0.12	0.35	1.46
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.31)
	simulation: add; estimation: rec								
FS1	900	0.24	0.17	0.26	0.14	0.34	-0.11	0.32	0.03
		(0.00)	(0.00)	(0.00)	(0.08)	(0.12)	(0.06)	(0.00)	(3.96)
FS3	600	0.24	0.17	0.26	0.23	0.42	-0.20	0.31	0.55
		(0.00)	(0.00)	(0.00)	(0.04)	(0.06)	(0.03)	(0.00)	(2.12)
FS4	600	0.24	0.17	0.26	0.21	0.48	-0.19	0.31	0.22
		(0.00)	(0.00)	(0.00)	(0.06)	(0.10)	(0.03)	(0.00)	(3.18)
FS6	450	0.24	0.17	0.26	0.23	0.40	-0.18	0.30	0.56
		(0.00)	(0.00)	(0.00)	(0.05)	(0.06)	(0.03)	(0.00)	(2.09)
	simulation: add; estimation: add								
FS1	900	0.24	0.17	0.26	0.18	0.35	-0.17	0.30	1.39
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.39)
FS3	600	0.24	0.17	0.26	0.15	0.29	-0.15	0.30	2.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)
FS4	600	0.24	0.17	0.26	0.16	0.32	-0.16	0.30	1.54
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.22)
FS6	450	0.24	0.17	0.26	0.15	0.28	-0.14	0.30	2.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)

Table C.32: Parameter estimates for a total of N = 1800 individuals and unconditional sampling. MSE is given in parentheses in a separate line. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$\eta_2 = 0.2$	$\eta_3 = 0.3$	$R_1 = 0.1$	$R_2 = 0.2$	$R_3 = -0.0$	$p^\star = 0.3$	$\beta = 2.0$
	simulation: dom; estimation: dom								
FS1	900	0.22	0.15	0.29	0.13	0.30	-0.03	0.32	1.33
		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)	(0.00)	(0.46)
FS3	600	0.22	0.15	0.29	0.09	0.24	-0.02	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.04)
FS4	600	0.22	0.15	0.29	0.11	0.28	-0.04	0.32	1.52
		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.01)	(0.00)	(0.25)
FS6	450	0.22	0.15	0.29	0.09	0.24	-0.02	0.31	1.98
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.04)
	simulation: dom; estimation: rec								
FS1	900	0.22	0.15	0.29	0.07	0.13	0.02	0.38	-0.05
		(0.00)	(0.00)	(0.00)	(0.06)	(0.15)	(0.07)	(0.02)	(4.25)
FS3	600	0.22	0.15	0.29	0.16	0.32	-0.04	0.32	0.65
		(0.00)	(0.00)	(0.00)	(0.07)	(0.06)	(0.04)	(0.00)	(1.86)
FS4	600	0.22	0.15	0.29	0.14	0.35	-0.03	0.32	0.21
		(0.00)	(0.00)	(0.00)	(0.05)	(0.09)	(0.04)	(0.01)	(3.24)
FS6	450	0.22	0.15	0.29	0.15	0.32	-0.03	0.31	0.67
		(0.00)	(0.00)	(0.00)	(0.06)	(0.07)	(0.04)	(0.00)	(1.81)
	simulation: dom; estimation: add								
FS1	900	0.22	0.15	0.29	0.13	0.30	-0.04	0.27	1.86
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.07)
FS3	600	0.22	0.15	0.29	0.09	0.24	-0.03	0.25	2.63
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.45)
FS4	600	0.22	0.15	0.29	0.11	0.25	-0.04	0.24	2.16
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.07)
FS6	450	0.22	0.15	0.29	0.09	0.24	-0.03	0.25	2.64
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.45)
	simulation: rec; estimation: dom								
FS1	900	0.22	0.15	0.29	0.08	0.18	-0.02	0.52	2.09
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.07)	(0.32)
FS3	600	0.22	0.15	0.29	0.07	0.17	-0.02	0.59	2.61
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.10)	(0.97)
FS4	600	0.22	0.15	0.29	0.08	0.19	-0.02	0.59	1.96
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.10)	(0.19)
FS6	450	0.22	0.15	0.29	0.07	0.18	-0.02	0.67	2.19
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.15)	(0.22)
	simulation: rec; estimation: rec								
FS1	900	0.22	0.15	0.29	0.20	0.31	-0.09	0.32	0.17
		(0.00)	(0.00)	(0.00)	(0.07)	(0.11)	(0.07)	(0.01)	(5.49)
FS3	600	0.22	0.15	0.29	0.09	0.23	-0.02	0.33	1.98

		(0.00)	(0.00)	(0.00)	(0.02)	(0.03)	(0.03)	(0.01)	(0.57)
FS4	600	0.22	0.15	0.29	0.18	0.35	-0.08	0.30	0.98
		(0.00)	(0.00)	(0.00)	(0.06)	(0.07)	(0.06)	(0.01)	(1.77)
FS6	450	0.22	0.15	0.29	0.10	0.24	-0.03	0.34	1.96
		(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.03)	(0.02)	(0.68)
	simulation: rec; estimation: add								
FS1	900	0.22	0.15	0.29	0.09	0.24	-0.03	0.49	2.76
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.04)	(0.97)
FS3	600	0.22	0.15	0.29	0.07	0.19	-0.02	0.49	4.43
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.04)	(9.17)
FS4	600	0.22	0.15	0.29	0.09	0.23	-0.03	0.53	2.66
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.06)	(0.60)
FS6	450	0.22	0.15	0.29	0.08	0.21	-0.02	0.57	3.12
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.08)	(1.56)
	simulation: add; estimation: dom								
FS1	900	0.22	0.15	0.29	0.09	0.26	-0.02	0.37	0.96
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(1.09)
FS3	600	0.22	0.15	0.29	0.07	0.18	-0.01	0.37	1.44
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.35)
FS4	600	0.22	0.15	0.29	0.09	0.22	-0.03	0.38	1.08
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.87)
FS6	450	0.22	0.15	0.29	0.08	0.19	-0.01	0.40	1.36
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.02)	(0.45)
	simulation: add; estimation: rec								
FS1	900	0.22	0.15	0.29	0.08	0.13	-0.00	0.38	-0.02
		(0.00)	(0.00)	(0.00)	(0.06)	(0.17)	(0.06)	(0.02)	(4.14)
FS3	600	0.22	0.15	0.29	0.13	0.34	-0.03	0.33	0.52
		(0.00)	(0.00)	(0.00)	(0.04)	(0.06)	(0.03)	(0.00)	(2.20)
FS4	600	0.22	0.15	0.29	0.14	0.37	-0.03	0.32	0.19
		(0.00)	(0.00)	(0.00)	(0.05)	(0.10)	(0.03)	(0.01)	(3.29)
FS6	450	0.22	0.15	0.29	0.13	0.34	-0.03	0.32	0.54
		(0.00)	(0.00)	(0.00)	(0.04)	(0.06)	(0.03)	(0.00)	(2.15)
	simulation: add; estimation: add								
FS1	900	0.22	0.15	0.29	0.11	0.29	-0.03	0.32	1.35
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.45)
FS3	600	0.22	0.15	0.29	0.09	0.23	-0.02	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.07)
FS4	600	0.22	0.15	0.29	0.10	0.26	-0.03	0.31	1.53
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.25)
FS6	450	0.22	0.15	0.29	0.09	0.23	-0.02	0.31	1.98
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)

Table C.33: Parameter estimates for a total of N = 1920 individuals and unconditional sampling. MSE is given in parentheses in a separate line. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.3$	$\eta_2 = 0.1$	$\eta_3 = 0.3$	$R_1 = 0.3$	$R_2 = 0.1$	$R_3 = -0.1$	$p^\star = 0.3$	$\beta = 2.0$
	simulation: dom; estimation: dom								
FS1a	960	0.29	0.12	0.26	0.37	0.12	-0.17	0.32	1.52
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.25)
FS3a	720	0.29	0.12	0.26	0.31	0.10	-0.14	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.03)
FS4a	720	0.29	0.12	0.26	0.35	0.11	-0.16	0.31	1.64
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.14)
FS6a	576	0.29	0.12	0.26	0.31	0.10	-0.14	0.31	1.98
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)
	simulation: dom; estimation: rec								
FS1a	960	0.29	0.12	0.26	0.45	0.21	-0.18	0.32	0.23
		(0.00)	(0.00)	(0.00)	(0.05)	(0.07)	(0.03)	(0.00)	(3.15)
FS3a	720	0.29	0.12	0.26	0.44	0.16	-0.19	0.31	0.66
		(0.00)	(0.00)	(0.00)	(0.04)	(0.04)	(0.02)	(0.00)	(1.81)
FS4a	720	0.29	0.12	0.26	0.48	0.20	-0.20	0.31	0.34
		(0.00)	(0.00)	(0.00)	(0.05)	(0.07)	(0.03)	(0.00)	(2.78)
FS6a	576	0.29	0.12	0.26	0.45	0.15	-0.20	0.31	0.66
		(0.00)	(0.00)	(0.00)	(0.05)	(0.04)	(0.02)	(0.00)	(1.80)
	simulation: dom; estimation: add								
FS1a	960	0.29	0.12	0.26	0.33	0.10	-0.16	0.23	2.20
		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.07)
FS3a	720	0.29	0.12	0.26	0.30	0.10	-0.14	0.25	2.66
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.46)
FS4a	720	0.29	0.12	0.26	0.30	0.10	-0.15	0.22	2.35
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.14)
FS6a	576	0.29	0.12	0.26	0.30	0.10	-0.14	0.24	2.66
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.45)
	simulation: rec; estimation: dom								
FS1a	960	0.29	0.12	0.26	0.28	0.09	-0.12	0.61	1.88
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.11)	(0.18)
FS3a	720	0.29	0.12	0.26	0.24	0.07	-0.12	0.63	2.36
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.13)	(0.46)
FS4a	720	0.29	0.12	0.26	0.27	0.09	-0.13	0.63	1.93
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.12)	(0.14)
FS6a	576	0.29	0.12	0.26	0.24	0.07	-0.13	0.65	2.24
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.14)	(0.26)
	simulation: rec; estimation: rec								
FS1a	960	0.29	0.12	0.26	0.50	0.17	-0.24	0.30	1.07
		(0.00)	(0.00)	(0.00)	(0.07)	(0.04)	(0.03)	(0.00)	(1.14)
FS3a	720	0.29	0.12	0.26	0.33	0.11	-0.15	0.33	1.96

		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)	(0.01)	(0.54)
FS4a	720	0.29	0.12	0.26	0.44	0.15	-0.21	0.29	1.43
		(0.00)	(0.00)	(0.00)	(0.05)	(0.03)	(0.03)	(0.00)	(0.79)
FS6a	576	0.29	0.12	0.26	0.32	0.11	-0.15	0.34	1.92
		(0.00)	(0.00)	(0.00)	(0.02)	(0.01)	(0.02)	(0.01)	(0.42)
	simulation: rec; estimation: add								
FS1a	960	0.28	0.12	0.26	0.31	0.10	-0.14	0.52	2.70
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.05)	(0.64)
FS3a	720	0.29	0.12	0.26	0.28	0.09	-0.13	0.57	3.13
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.08)	(1.47)
FS4a	720	0.29	0.12	0.26	0.30	0.10	-0.14	0.53	2.74
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)	(0.65)
FS6a	576	0.29	0.12	0.26	0.27	0.09	-0.13	0.58	2.99
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.08)	(1.07)
	simulation: add; estimation: dom								
FS1a	960	0.29	0.12	0.26	0.28	0.09	-0.13	0.34	1.14
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.75)
FS3a	720	0.29	0.12	0.26	0.26	0.08	-0.12	0.40	1.36
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.02)	(0.44)
FS4a	720	0.29	0.12	0.26	0.28	0.09	-0.13	0.35	1.19
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.67)
FS6a	576	0.29	0.12	0.26	0.27	0.08	-0.12	0.41	1.33
		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.02)	(0.48)
	simulation: add; estimation: rec								
FS1a	960	0.29	0.12	0.26	0.50	0.18	-0.20	0.32	0.22
		(0.00)	(0.00)	(0.00)	(0.09)	(0.07)	(0.03)	(0.00)	(3.19)
FS3a	720	0.29	0.12	0.26	0.46	0.16	-0.19	0.32	0.54
		(0.00)	(0.00)	(0.00)	(0.05)	(0.04)	(0.02)	(0.00)	(2.15)
FS4a	720	0.29	0.12	0.26	0.54	0.17	-0.23	0.31	0.30
		(0.00)	(0.00)	(0.00)	(0.09)	(0.06)	(0.03)	(0.00)	(2.91)
FS6a	576	0.29	0.12	0.26	0.46	0.16	-0.19	0.32	0.54
		(0.00)	(0.00)	(0.00)	(0.05)	(0.04)	(0.02)	(0.00)	(2.16)
	simulation: add; estimation: add								
FS1a	960	0.29	0.12	0.26	0.34	0.11	-0.16	0.31	1.52
		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.25)
FS3a	720	0.29	0.12	0.26	0.31	0.10	-0.14	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)
FS4a	720	0.29	0.12	0.26	0.32	0.10	-0.15	0.30	1.65
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)
FS6a	576	0.29	0.12	0.26	0.30	0.10	-0.14	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)

Table C.34: Parameter estimates for a total of N = 1920 individuals and unconditional sampling. MSE is given in parentheses in a separate line. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$\eta_2 = 0.2$	$\eta_3 = 0.3$	$R_1 = 0.1$	$R_2 = 0.3$	$R_3 = -0.1$	$p^\star = 0.3$	$\beta = 2.0$
	simulation: dom; estimation: dom								
FS1a	960	0.24	0.17	0.26	0.17	0.35	-0.16	0.31	1.54
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.22)
FS3a	720	0.24	0.17	0.26	0.15	0.29	-0.14	0.30	2.00
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)
FS4a	720	0.24	0.17	0.26	0.16	0.32	-0.15	0.30	1.64
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.14)
FS6a	576	0.24	0.17	0.26	0.15	0.29	-0.14	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)
	simulation: dom; estimation: rec								
FS1a	960	0.24	0.17	0.26	0.24	0.45	-0.18	0.31	0.24
		(0.00)	(0.00)	(0.00)	(0.06)	(0.08)	(0.03)	(0.00)	(3.13)
FS3a	720	0.24	0.17	0.26	0.23	0.39	-0.19	0.30	0.68
		(0.00)	(0.00)	(0.00)	(0.05)	(0.05)	(0.03)	(0.00)	(1.76)
FS4a	720	0.24	0.17	0.26	0.25	0.46	-0.20	0.30	0.34
		(0.00)	(0.00)	(0.00)	(0.06)	(0.08)	(0.03)	(0.00)	(2.76)
FS6a	576	0.24	0.17	0.26	0.22	0.40	-0.18	0.30	0.68
		(0.00)	(0.00)	(0.00)	(0.04)	(0.05)	(0.03)	(0.00)	(1.76)
	simulation: dom; estimation: add								
FS1a	960	0.24	0.17	0.26	0.16	0.33	-0.16	0.25	2.12
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.05)
FS3a	720	0.24	0.17	0.26	0.15	0.29	-0.14	0.25	2.63
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.42)
FS4a	720	0.24	0.17	0.26	0.15	0.30	-0.15	0.23	2.29
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.11)
FS6a	576	0.24	0.17	0.26	0.14	0.29	-0.14	0.25	2.63
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.42)
	simulation: rec; estimation: dom								
FS1a	960	0.24	0.17	0.26	0.12	0.23	-0.12	0.55	2.15
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.08)	(0.30)
FS3a	720	0.24	0.17	0.26	0.11	0.20	-0.11	0.61	2.46
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.11)	(0.58)
FS4a	720	0.24	0.17	0.26	0.12	0.23	-0.12	0.59	2.08
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.10)	(0.21)
FS6a	576	0.24	0.17	0.26	0.11	0.22	-0.12	0.64	2.27
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.13)	(0.26)
	simulation: rec; estimation: rec								
FS1a	960	0.24	0.17	0.26	0.24	0.47	-0.24	0.30	1.03
		(0.00)	(0.00)	(0.00)	(0.05)	(0.07)	(0.04)	(0.00)	(1.09)
FS3a	720	0.24	0.17	0.26	0.15	0.29	-0.15	0.31	1.98

		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.01)	(0.00)	(0.25)
FS4a	720	0.24	0.17	0.26	0.22	0.42	-0.21	0.29	1.34
		(0.00)	(0.00)	(0.00)	(0.03)	(0.05)	(0.03)	(0.00)	(0.65)
FS6a	576	0.24	0.17	0.26	0.15	0.29	-0.14	0.31	1.98
		(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.01)	(0.00)	(0.22)
	simulation: rec; estimation: add								
FS1a	960	0.24	0.17	0.26	0.14	0.28	-0.14	0.48	2.93
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(1.15)
FS3a	720	0.24	0.17	0.26	0.13	0.25	-0.13	0.54	3.40
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.06)	(2.59)
FS4a	720	0.24	0.17	0.26	0.15	0.27	-0.14	0.52	2.79
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.05)	(0.74)
FS6a	576	0.24	0.17	0.26	0.13	0.26	-0.13	0.57	3.09
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.08)	(1.36)
	simulation: add; estimation: dom								
FS1a	960	0.24	0.17	0.26	0.13	0.27	-0.13	0.33	1.17
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.71)
FS3a	720	0.24	0.17	0.26	0.12	0.22	-0.12	0.35	1.47
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.30)
FS4a	720	0.24	0.17	0.26	0.13	0.25	-0.13	0.34	1.21
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.64)
FS6a	576	0.24	0.17	0.26	0.12	0.23	-0.12	0.36	1.43
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.35)
	simulation: add; estimation: rec								
FS1a	960	0.24	0.17	0.26	0.18	0.53	-0.19	0.31	0.22
		(0.00)	(0.00)	(0.00)	(0.04)	(0.10)	(0.02)	(0.00)	(3.17)
FS3a	720	0.24	0.17	0.26	0.22	0.45	-0.20	0.31	0.55
		(0.00)	(0.00)	(0.00)	(0.03)	(0.06)	(0.02)	(0.00)	(2.12)
FS4a	720	0.24	0.17	0.26	0.22	0.53	-0.21	0.31	0.30
		(0.00)	(0.00)	(0.00)	(0.05)	(0.10)	(0.02)	(0.00)	(2.89)
FS6a	576	0.24	0.17	0.26	0.22	0.44	-0.19	0.30	0.56
		(0.00)	(0.00)	(0.00)	(0.03)	(0.05)	(0.02)	(0.00)	(2.09)
	simulation: add; estimation: add								
FS1a	960	0.24	0.17	0.26	0.16	0.32	-0.16	0.30	1.54
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.22)
FS3a	720	0.24	0.17	0.26	0.15	0.29	-0.14	0.30	2.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)
FS4a	720	0.24	0.17	0.26	0.15	0.31	-0.15	0.30	1.64
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)
FS6a	576	0.24	0.17	0.26	0.14	0.29	-0.14	0.30	2.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)

Table C.35: Parameter estimates for a total of N = 1920 individuals and unconditional sampling. MSE is given in parentheses in a separate line. n number of families; $\mu = -3.00$; number of iterations 1000.

-	n	$\eta_1 = 0.2$	$\eta_2 = 0.2$	$\eta_3 = 0.3$	$R_1 = 0.1$	$R_2 = 0.2$	$R_3 = -0.0$	$p^\star = 0.3$	$\beta = 2.0$
	simulation: dom; estimation: dom								
FS1a	960	0.22	0.15	0.29	0.11	0.28	-0.03	0.32	1.51
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.25)
FS3a	720	0.22	0.15	0.29	0.09	0.23	-0.02	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.02)
FS4a	720	0.22	0.15	0.29	0.10	0.27	-0.03	0.31	1.63
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.15)
FS6a	576	0.22	0.15	0.29	0.08	0.24	-0.02	0.31	2.00
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.02)
	simulation: dom; estimation: rec								
FS1a	960	0.22	0.15	0.29	0.15	0.37	-0.03	0.31	0.21
		(0.00)	(0.00)	(0.00)	(0.04)	(0.09)	(0.03)	(0.00)	(3.24)
FS3a	720	0.22	0.15	0.29	0.15	0.35	-0.03	0.31	0.66
		(0.00)	(0.00)	(0.00)	(0.04)	(0.05)	(0.03)	(0.00)	(1.82)
FS4a	720	0.22	0.15	0.29	0.14	0.42	-0.04	0.30	0.32
		(0.00)	(0.00)	(0.00)	(0.04)	(0.08)	(0.03)	(0.00)	(2.83)
FS6a	576	0.22	0.15	0.29	0.14	0.34	-0.04	0.30	0.68
		(0.00)	(0.00)	(0.00)	(0.04)	(0.05)	(0.03)	(0.00)	(1.77)
	simulation: dom; estimation: add								
FS1a	960	0.22	0.15	0.29	0.11	0.25	-0.03	0.24	2.15
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.06)
FS3a	720	0.22	0.15	0.29	0.09	0.24	-0.03	0.25	2.65
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.45)
FS4a	720	0.22	0.15	0.29	0.09	0.23	-0.03	0.22	2.33
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.13)
FS6a	576	0.22	0.15	0.29	0.09	0.23	-0.03	0.25	2.64
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.43)
	simulation: rec; estimation: dom								
FS1a	960	0.22	0.15	0.29	0.08	0.20	-0.02	0.60	1.93
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.10)	(0.18)
FS3a	720	0.22	0.15	0.29	0.07	0.19	-0.02	0.68	2.15
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.15)	(0.14)
FS4a	720	0.22	0.15	0.29	0.07	0.20	-0.03	0.61	1.98
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.11)	(0.14)
FS6a	576	0.22	0.15	0.29	0.07	0.19	-0.02	0.69	2.09
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.16)	(0.09)
	simulation: rec; estimation: rec								
FS1a	960	0.22	0.15	0.29	0.18	0.39	-0.08	0.29	1.01
		(0.00)	(0.00)	(0.00)	(0.05)	(0.07)	(0.04)	(0.00)	(1.35)
FS3a	720	0.22	0.15	0.29	0.10	0.24	-0.02	0.34	1.93

		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)	(0.01)	(0.48)
FS4a	720	0.22	0.15	0.29	0.15	0.35	-0.07	0.28	1.39
		(0.00)	(0.00)	(0.00)	(0.04)	(0.05)	(0.03)	(0.01)	(0.80)
FS6a	576	0.22	0.15	0.29	0.10	0.24	-0.03	0.34	1.90
		(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.01)	(0.01)	(0.46)
	simulation: rec; estimation: add								
FS1a	960	0.22	0.15	0.29	0.09	0.24	-0.03	0.53	2.64
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)	(0.54)
FS3a	720	0.22	0.15	0.29	0.08	0.21	-0.03	0.57	3.10
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.08)	(1.38)
FS4a	720	0.22	0.15	0.29	0.09	0.22	-0.02	0.55	2.66
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)	(0.51)
FS6a	576	0.22	0.15	0.29	0.09	0.21	-0.02	0.59	2.93
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.09)	(0.92)
	simulation: add; estimation: dom								
FS1a	960	0.22	0.15	0.29	0.08	0.23	-0.02	0.38	1.07
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.89)
FS3a	720	0.22	0.15	0.29	0.08	0.18	-0.02	0.39	1.37
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.42)
FS4a	720	0.22	0.15	0.29	0.08	0.22	-0.03	0.38	1.14
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.75)
FS6a	576	0.22	0.15	0.29	0.08	0.19	-0.02	0.41	1.34
		(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.02)	(0.47)
	simulation: add; estimation: rec								
FS1a	960	0.22	0.15	0.29	0.14	0.41	-0.04	0.31	0.19
		(0.00)	(0.00)	(0.00)	(0.04)	(0.10)	(0.02)	(0.00)	(3.28)
FS3a	720	0.22	0.15	0.29	0.13	0.37	-0.03	0.32	0.53
		(0.00)	(0.00)	(0.00)	(0.03)	(0.05)	(0.02)	(0.00)	(2.17)
FS4a	720	0.22	0.15	0.29	0.14	0.44	-0.04	0.31	0.28
		(0.00)	(0.00)	(0.00)	(0.04)	(0.09)	(0.03)	(0.00)	(2.96)
FS6a	576	0.22	0.15	0.29	0.13	0.36	-0.04	0.32	0.54
		(0.00)	(0.00)	(0.00)	(0.03)	(0.05)	(0.02)	(0.00)	(2.15)
	simulation: add; estimation: add								
FS1a	960	0.22	0.15	0.29	0.10	0.26	-0.03	0.31	1.52
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.25)
FS3a	720	0.22	0.15	0.29	0.09	0.23	-0.02	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)
FS4a	720	0.22	0.15	0.29	0.09	0.24	-0.02	0.30	1.64
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.15)
FS6a	576	0.22	0.15	0.29	0.09	0.23	-0.02	0.31	1.99
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.03)

C.5. **MCMC convergence plots.** The following figures show examples for MCMC runs. The burnin samples are not discarded in order to allow judgement of convergence speed. The first two chains show convergence whereas the last chain does not show convergence for a misspecified model.

Figure C.1: Display of the markov chain of the MCMC sampler. Top left is the original chain $(\eta_1^*, \eta_2^*, \eta_3^*, \eta_4^*, \beta) = (.1, .3, .15, .45, 2)$. Top right shows the reparametrized chain in terms of parametrization 4.The bottom diplays the prior (thin line) and a density estimate of the posterior distributions (thick line).

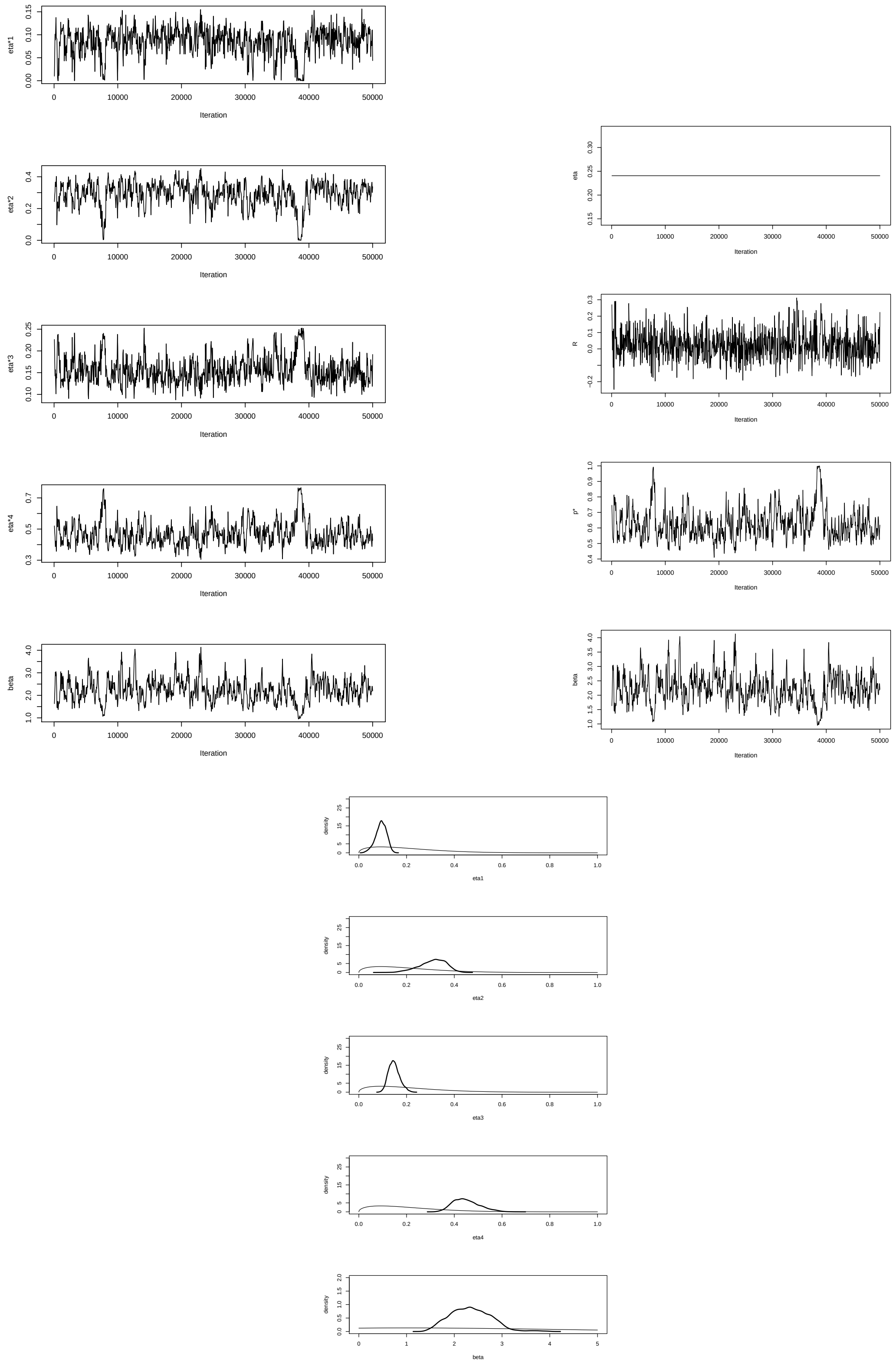


Figure C.2: Display of the markov chain of the MCMC sampler. Top left is the original chain $(\eta_1^*, \eta_2^*, \eta_3^*, \eta_4^*, \beta) = (.01, .78, .2, .01, 2.5)$. Top right shows the reparametrized chain in terms of parametrization 4.The bottom diplays the prior (thin line) and a density estimate of the posterior distributions (thick line).

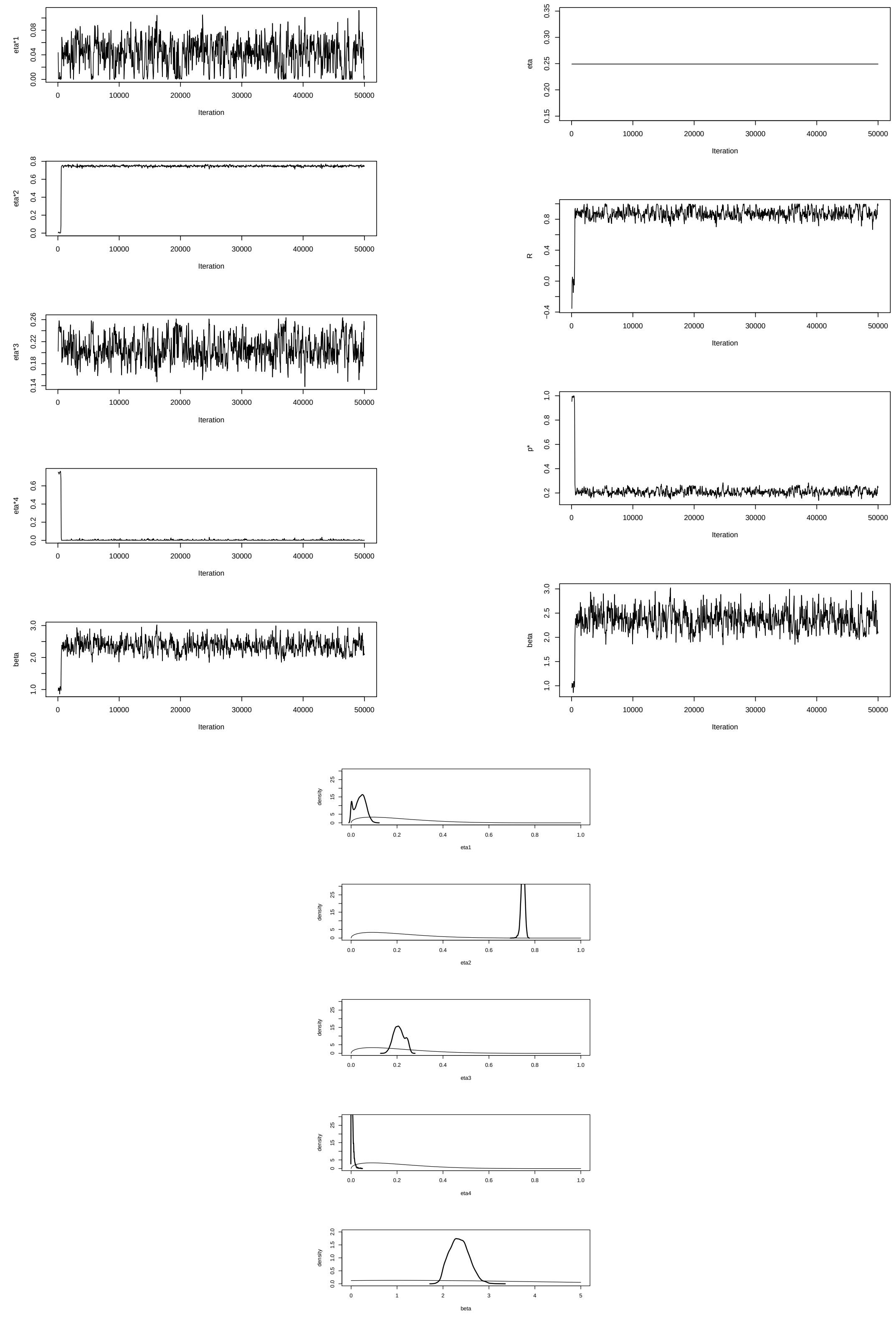


Figure C.3: Display of the markov chain of the MCMC sampler. Top left is the original chain $(\eta_1^*, \eta_2^*, \eta_3^*, \eta_4^*, \beta) = (.125, .5, .075, .3, 2)$. Top right shows the reparametrized chain in terms of parametrization 4.The bottom diplays the prior (thin line) and a density estimate of the posterior distributions (thick line). Simulation: recessive, Estimation: additive.

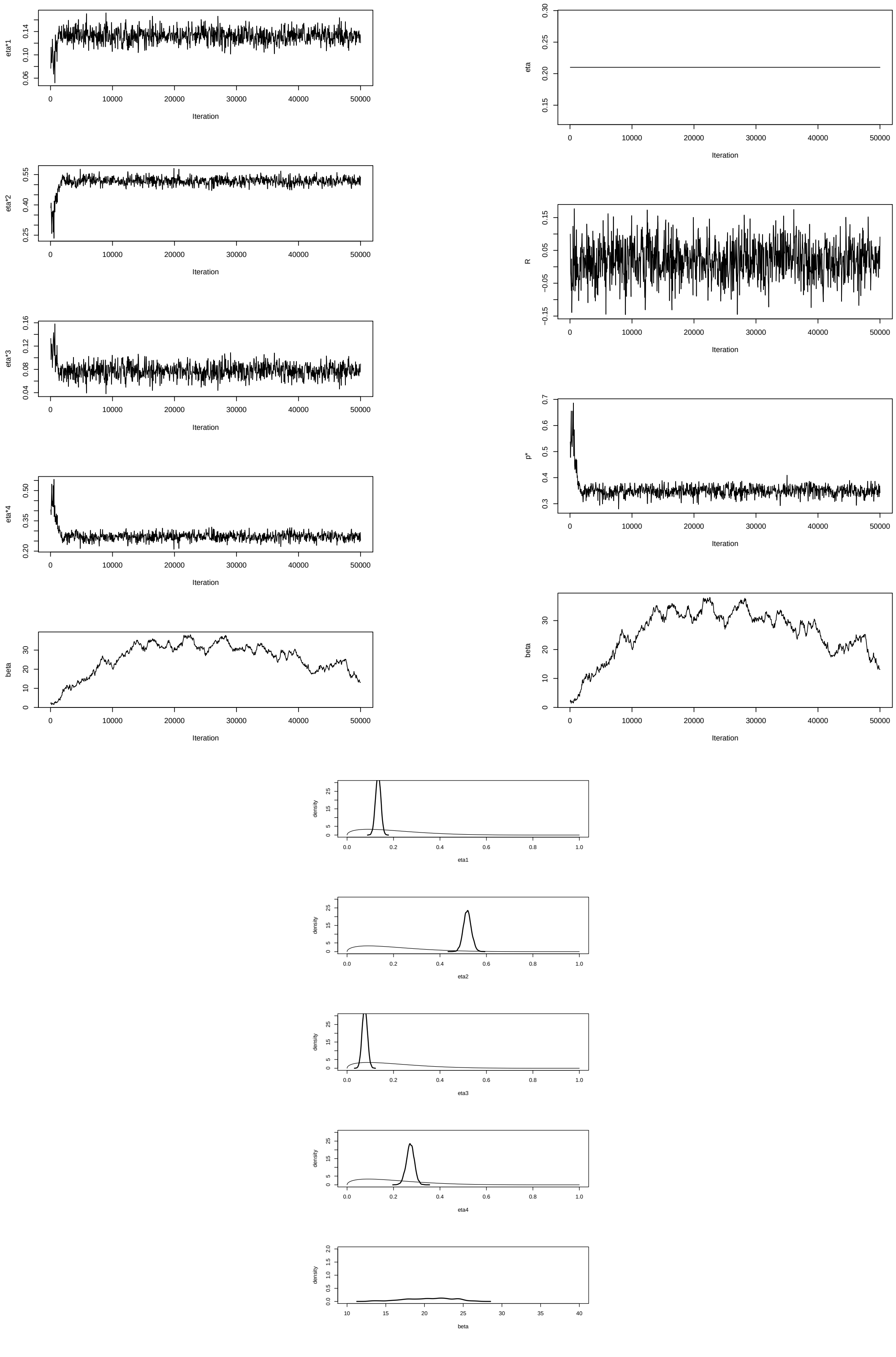


Table D.1: Data analysis for a single SNP analysis of the Alzheimer's data set assuming recessive inheritance. *LR* is the likelihood ratio statistics and p-value is based on a χ_1^2 approximation. Parameter estimates are given under the alternative hypothesis.

	SNP	model	LR	p-value	η_1	R^2	$p^\star$	β
3	SNP1006	rec	-6.734e+02	9.958e-01	0.66	0.00	1.00	1.19
6	SNP875	rec	-6.678e+02	9.944e-01	0.70	0.00	1.00	1.19
9	SNP886	rec	-6.588e+02	9.982e-01	0.71	0.00	1.00	1.19
12	SNP988	rec	-6.805e+02	1.845e-01	0.60	-0.53	0.70	1.57
15	SNP888	rec	-6.642e+02	5.359e-02	0.29	-0.55	0.88	1.34
18	SNP873	rec	-6.554e+02	**3.230e-02**	0.72	0.57	0.89	1.34
21	SNP952	rec	-6.692e+02	9.966e-01	0.72	0.00	1.00	1.19
24	SNP528	rec	-6.855e+02	**6.086e-03**	0.39	0.46	0.75	1.58
27	SNP992	rec	-6.464e+02	9.964e-01	0.27	-0.00	1.00	1.19
30	SNP465	rec	-7.072e+02	9.980e-01	0.44	0.00	1.00	1.19
33	SNP457	rec	-7.202e+02	9.982e-01	0.51	0.00	1.00	1.19
36	SNP471	rec	-7.171e+02	9.987e-01	0.54	0.00	1.00	1.19
39	SNP479	rec	-6.199e+02	9.982e-01	0.33	0.00	1.00	1.19
42	SNP497	rec	-7.080e+02	9.982e-01	0.56	0.00	1.00	1.19
45	SNP491	rec	-4.046e+02	7.966e-01	0.99	0.14	1.00	1.20
48	SNP459	rec	-6.893e+02	9.995e-01	0.50	0.00	1.00	1.19
51	SNP512	rec	-6.089e+02	9.979e-01	0.78	0.00	1.00	1.19

Table D.2: Data analysis for a single SNP analysis of the Alzheimer's data set assuming recessive inheritance and a random effect. l_A is the supremum of the likelihood under the alternative; p-value is based on a χ_1^2 approximation. Parameter estimates are given under the alternative hypothesis.

	SNP	model	LR	p-value	η_1	R^2	$p^\star$	β	σ
3	SNP1006	rec	-673.43	0.99	0.65	0	1	1.19	2e-5
6	SNP875	rec	-667.83	0.87	0.69	2e-3	0.99	1.19	0
9	SNP886	rec	-658.75	1.00	0.71	0	1	1.19	0
12	SNP988	rec	-679.53	0.05	0.58	-0.41	0.81	1.46	0
15	SNP888	rec	-664.15	0.05	0.29	-0.55	0.89	1.32	4e-4
18	SNP873	rec	-655.40	**0.03**	0.72	0.57	0.89	1.34	2e-4
21	SNP952	rec	-669.21	0.92	0.72	0.13	0.99	1.20	0
24	SNP528	rec	-685.53	**0.01**	0.39	0.47	0.75	1.59	0.002
27	SNP992	rec	-646.44	1.00	0.27	-9.13e-05	1	1.19	7e-6
30	SNP465	rec	-707.15	1.00	0.44	0	1	1.19	0
33	SNP457	rec	-720.23	1.00	0.51	0	1	1.19	0
36	SNP471	rec	-717.07	1.00	0.54	4.55e-05	1	1.19	4e-5
39	SNP479	rec	-619.87	1.00	0.33	0	1	1.19	0
42	SNP497	rec	-708.02	1.00	0.56	2.32e-05	1	1.19	0
45	SNP491	rec	-404.55	0.80	0.99	0.06	1.00	1.19	0
48	SNP459	rec	-689.30	1.00	0.50	0	1	1.19	0
51	SNP512	rec	-608.90	1.00	0.78	1.39e-05	1	1.19	0

Table D.3: Data analysis for haplotypes of the Alzheimer's data set. The p-value based on a χ^2_3 approximation of the LR-statistic. Parameter estimates are given under the alternative hypothesis. SNP is abbreviated as S, in order to make the table more legible (S1006:S875 is the model for SNPs SNP1006, SNP875).

	SNP	model	p-value	η_1	η_2	η_3	R^2_1	R^2_2	R^2_3	$p^\star$	β
1	S1006:S875	add	8.3e-01	0.345	0.343	0.312	0.000	0.000	0.000	0.431	1.010
2	S1006:S875	dom	8.1e-01	0.364	0.336	0.300	0.000	0.000	0.000	0.402	1.468
3	S1006:S875	rec	3.4e-01	0.358	0.345	0.297	-0.075	-0.055	0.136	0.496	2.165
4	S875:S886	add	8.2e-01	0.650	0.060	0.048	-0.056	-0.030	0.026	0.987	0.600
5	S875:S886	dom	8.1e-01	0.656	0.062	0.046	-0.331	-0.147	0.157	0.662	1.291
6	S875:S886	rec	1.0e-00	0.648	0.060	0.049	-0.000	0.000	-0.000	1.000	1.189
7	S886:S988	add	**4.0e-02**	0.401	0.183	0.340	-0.677	0.170	0.429	0.593	0.834
8	S886:S988	dom	**9.1e-04**	0.375	0.228	0.320	-0.559	-0.254	0.768	0.444	1.462
9	S886:S988	rec	1.8e-01	0.415	0.179	0.327	-0.625	0.192	0.402	0.749	1.483
10	S988:S888	add	1.1e-01	0.304	0.028	0.307	-0.654	0.131	0.030	0.623	0.765
11	S988:S888	dom	**6.8e-04**	0.254	0.027	0.314	-0.384	-0.155	-0.378	0.468	1.434
12	S988:S888	rec	2.0e-01	0.265	0.031	0.300	-0.551	-0.071	0.232	0.880	1.337
13	S888:S873	add	2.9e-01	0.047	0.767	0.186	-0.000	-0.000	0.000	0.296	1.281
14	S888:S873	dom	**4.3e-08**	0.103	0.686	0.211	-0.234	0.490	-0.383	0.718	1.264
15	S888:S873	rec	4.7e-01	0.037	0.673	0.289	-0.000	0.000	-0.000	0.270	23.440
16	S873:S952	add	1.2e-01	0.426	0.236	0.339	0.340	-0.363	-0.030	0.494	0.906
17	S873:S952	dom	**1.2e-03**	0.389	0.253	0.358	0.702	-0.358	-0.389	0.420	1.432
18	S873:S952	rec	6.8e-01	0.478	0.266	0.255	0.115	-0.174	0.058	0.982	1.225
19	S952:S528	add	1.0e-00	0.429	0.004	0.289	0.015	-0.000	-0.000	1.000	0.595
20	S952:S528	dom	**6.8e-05**	0.408	0.007	0.315	0.791	-0.025	-0.592	0.521	1.441
21	S952:S528	rec	**4.2e-02**	0.361	0.003	0.362	0.526	-0.074	-0.569	0.672	1.705
22	S528:S992	add	1.0e-00	0.004	0.270	0.429	-0.000	0.001	0.014	1.000	0.595
23	S528:S992	dom	**1.6e-05**	0.008	0.265	0.385	-0.081	-0.297	0.901	0.436	1.392
24	S528:S992	rec	**2.2e-03**	0.008	0.264	0.399	0.043	-0.004	0.400	0.805	1.493
25	S992:S465	add	1.0e-00	0.150	0.293	0.121	0.000	0.000	0.000	1.000	0.595
26	S992:S465	dom	3.9e-01	0.163	0.291	0.117	-0.624	0.070	-0.215	0.719	1.269
27	S992:S465	rec	1.0e-00	0.150	0.293	0.121	0.000	-0.000	-0.000	1.000	1.189
28	S465:S457	add	8.8e-01	0.287	0.238	0.185	-0.010	-0.148	-0.408	0.621	0.797
29	S465:S457	dom	3.2e-01	0.271	0.235	0.179	0.309	-0.142	-0.542	0.577	1.334
30	S465:S457	rec	7.5e-01	0.257	0.246	0.194	0.245	-0.146	-0.430	0.852	1.371
31	S457:S471	add	1.0e-00	0.505	0.028	0.001	0.016	0.006	-0.000	0.999	0.596
32	S457:S471	dom	9.5e-01	0.503	0.036	0.001	0.005	-0.038	-0.002	1.000	1.198
33	S457:S471	rec	8.2e-01	0.498	0.030	0.003	0.017	-0.006	-0.001	1.000	1.198
34	S471:S479	add	1.0e-00	0.153	0.169	0.387	-0.000	0.000	-0.001	1.000	0.604
35	S471:S479	dom	5.5e-01	0.150	0.161	0.405	0.063	0.337	-0.427	0.609	1.343
36	S471:S479	rec	1.0e-00	0.153	0.169	0.387	-0.000	0.000	-0.001	1.000	1.208
37	S479:S497	add	1.0e-00	0.189	0.384	0.137	0.000	0.001	0.000	1.000	0.572
38	S479:S497	dom	8.0e-01	0.176	0.378	0.155	0.356	-0.166	-0.251	0.606	1.227
39	S479:S497	rec	1.0e-00	0.188	0.384	0.139	0.000	0.001	-0.000	1.000	1.144

40	S497:S491	add	6.9e-01	0.561	0.429	0.000	0.223	-0.130	-0.023	0.960	0.612
41	S497:S491	dom	1.4e-01	0.603	0.387	0.000	0.013	0.007	-0.000	0.510	1.386
42	S497:S491	rec	7.8e-01	0.542	0.449	0.000	0.112	-0.104	0.002	0.278	54.083
43	S491:S459	add	2.3e-01	0.500	0.002	0.487	-0.040	-0.104	0.093	0.818	0.664
44	S491:S459	dom	9.9e-01	0.497	0.005	0.491	0.013	-0.018	-0.010	1.000	1.186
45	S491:S459	rec	9.9e-01	0.497	0.004	0.491	0.013	-0.006	0.015	1.000	1.189
46	S459:S512	add	1.0e-00	0.320	0.453	0.184	0.000	-0.000	-0.000	1.000	0.594
47	S459:S512	dom	1.3e-01	0.321	0.432	0.206	-0.506	0.685	-0.335	0.614	1.355
48	S459:S512	rec	1.0e-00	0.320	0.453	0.183	-0.001	0.000	0.001	1.000	1.189

References

[1] Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 3:299–309.

[2] Armitage P, Colton T, eds. (1998) *Encyclopedia of Biostatistics.* Wiley.

[3] Bacanu SA, Devlin B, Roeder K (2000) The Power of Genomic Control. *Am J Hum Genet* 66:1933–1944.

[4] Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7:781–791.

[5] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B* 57:289–300.

[6] Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol* 2:85–97.

[7] Boehringer S, Hardt C, Miterski B, Steland A, Epplen JT (2003) Multilocus statistics to uncover epistasis and heterogeneity in complex, diseases: revisiting a set of multiple sclerosis data. *Eur J Hum Genet* 11:573–84.

[8] Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. *Nat Genet* 37:868–872.

[9] Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 74:106–120.

[10] Chen WM, Abecasis GR (2007) Family-based association tests for genomewide association scans. *Am J Hum Genet* 81:913–926.

[11] Chiou JM, Liang KY, Chiu YF (2005) Multipoint linkage mapping using sibpairs: non-parametric estimation of, trait effects with quantitative covariates. *Genet Epidemiol* 28:58–69.

[12] Chung RH, Morris RW, Zhang L, Li YJ, Martin ER (2007) X-APL: an improved family-based test of association in the presence of linkage for the X chromosome. *Am J Hum Genet* 80:59–68.

[13] Consortium IH (2003) The International HapMap Project. *Nature* 426:789–96.

[14] Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.

[15] Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261:921–923.

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

[17] Edwards AO, Ritter R, Abel KJ, Manning A, Panhuysen C, Farrer LA (2005) Complement factor H polymorphism and age-related macular degeneration. *Science* 308:421–424.

[18] Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap.* Chapman & Hall.

[19] Eichner JE, Dunn ST, Perveen G, Thompson DM, Stewart KE, Stroehla BC (2002) Apolipoprotein E polymorphism and cardiovascular disease: a HuGE review. *Am J Epidemiol* 155:487–95.

[20] Elston RC, Song D, Iyengar SK (2005) Mathematical assumptions versus biological reality: myths in affected sib pair linkage analysis. *Am J Hum Genet* 76:152–156.

[21] Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a, diploid population. *Mol Biol Evol* 12:921–7.

[22] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296:2225–2229.

[23] Gilks WR, Richardson S, Spiegelhalter DJ, eds. (1996) *Markov Chain Monte Carlo in Practice.* Chapman & Hall.

[24] Göring HH, Terwilliger JD (2000) Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet* 66:1095–1106.

[25] Göring HH, Terwilliger JD (2000) Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am J Hum Genet* 66:1107–1118.

[26] Göring HH, Terwilliger JD (2000) Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. *Am J Hum Genet* 66:1298–1309.

[27] Göring HH, Terwilliger JD (2000) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* 66:1310–1327.

[28] Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA (2006) Centralizing the non-central chi-square: A new method to correct for population stratification in genetic case-control association studies. *Genet Epidemiol* 30:277–289.

[29] Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA (2005) Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308:419–421.

[30] Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, et al. (2006) A common genetic variant is associated with adult and childhood obesity. *Science* 312:279–283.

[31] Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6:95–108.

[32] Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–802.

[33] Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. *Stat Med* 9:811–818.

[34] Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association, studies. *Genome Res* 11:2115–9.

[35] Holm S (1979) A simple sequentially rejective test procedure. *Scand J. Statist.* 6:65–70.

[36] Hommel G (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75:383–386.

[37] Horvath S, Xu X, Laird NM (2001) The family based association test method: strategies for studying general genotype–phenotype associations. *Eur J Hum Genet* 9:301–306.

[38] Horvath S KM Laird NM (2000) The transmission/disequilibrium test and parental-genotype reconstruction, for X-chromosomal markers. *Am J Hum Genet* 66:1161–7.

[39] Huang BE, Amos CI, Lin DY (2007) Detecting haplotype effects in genomewide association studies. *Genet Epidemiol* 31:803–812.

[40] Hudson RR (1990) Gene genealogies and the coalescent process. *Oxf Surv Evol Biol* 7:1–44.

[41] International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–1320.

[42] Kauermann G, Carroll R (2001) A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 96:1387–1396.

[43] Kazazian HH (2004) Mobile Elements: Drivers of Genome Evolution Mobile Elements: Drivers of Genome Evolution. *Science* 303:1626–32.

[44] Kimmel G, Shamir R (2006) A fast method for computing high-significance disease association in large population-based studies. *Am J Hum Genet* 79:481–492.

[45] Kingman JF (2000) Origins of the coalescent. 1974-1982. *Genetics* 156:1461–1463.

[46] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385–389.

[47] Knapp M (2001) Reconstructing parental genotypes when testing for linkage in the presence of association. *Theor Popul Biol* 60:141–148.

[48] Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 22:139–144.

[49] Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am J Hum Genet* 56:519–27.

[50] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 58:1347–63.

[51] Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 7:385–394.

[52] Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* 84:2363–2367.

[53] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

[54] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

[55] Lee W (2002) Transmission/disequilibrium test when neither parent is available in some families: a non-iterative approach. *J Cancer Epidemiol Prev* 7:97–103.

[56] Lehmann EL (1986) *Elements of Large-Sample Theory.* Springer, New York.

[57] Lewin B (1994) *Genes V.* Oxford University Press.

[58] Lewontin RC (1964) The interaction of selection and linkage. II. Optimum models. *Genetics* 50:757–782.

[59] Lin DY, Zeng D (2006) Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies. *JASA* 101:89–104.

[60] Lipshutz RF, et al (1994) Advanced DNA sequencing technologies. *Curr Opin Struct Biol* 4:376–380.

[61] Liu SL (1994) The collapsed Gibbs sampler in Baysian computations with applications to a gene regulation problem. *J Am Stat Assoc* 89:958–966.

[62] Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P, Consortium IH (2006) A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 78:437–450.

[63] Martin E, Bass M, Hauser E, Kaplan N (2003) Accounting for linkage in family-based tests of association with missing, parental genotypes. *Am J Hum Genet* 73:1016–26.

[64] Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, et al. (2000) SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am J Hum Genet* 67:383–94.

[65] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9:356–369.

[66] McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.

[67] Narain P (1990) *Statistical Genetics.* Wiley Eastern Limited.

[68] National Library of Medicine (1992) Online Mendelian Inheritance in Man. `http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim`, retrieved 12/2007.

[69] National Library of Medicine (1992) Online Mendelian Inheritance in Man. Alzheimer's disease. `http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=104300`, retrieved 12/2007.

[70] National Library of Medicine (2005) Genetic Home Reference. `http://ghr.nlm.nih.gov/condition=alzheimerdisease`, retrieved 12/2007.

[71] Nelson MR, Kardia SL, Ferrell RE, Sing CF (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 11:458–470.

[72] Niu T (2004) Algorithms for inferring haplotypes. *Genet Epidemiol* 27:334–47.

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

[74] Oksenberg JR, Baranzini SE, Sawcer S, Hauser SL (2008) The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. *Nat Rev Genet* 9:516–526.

[75] Oliphant A, Barker D, Stuelpnagel JR, Chee MS (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* Suppl:56-8:60–1.

[76] Pericak-Vance MA, Bebout JL, Gaskell PC, Yamaoka LH, Hung WY, Alberts MJ, Walker AP, Bartlett RJ, Haynes CA, Welsh KA (1991) Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am J Hum Genet* 48:1034–1050.

[77] Pfeiffer RM, Gail MH, Pee D (2001) Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika* 88:933–948.

[78] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical recipes in C.* Cambridge University Press.

[79] Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–59.

[80] Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *Am J Hum Genet* 67:170–81.

[81] Rabinowitz D (2003) Adjusting for population heterogeneity: A framework for characterizing statistical information and developing efficient test statistics. *Genet Epidemiol* 24:284–90.

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

[83] Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 20:4–16.

[84] Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–7.

[85] Risch N, 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* 8:1273–1288.

[86] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69:138–147.

[87] Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk C, Ginsberg F (1981) Genetics of HLA disease associations. The use of the haplotype relative risk (HRR) and the "haplo-delta" estimates in juvenile diabetes from three racial groups. *Human Immunology* 3:384.

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

[89] Seaman SR, Müller-Myhsok B (2005) Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. *Am J Hum Genet* 76:399–408.

[90] Smith CA (1975) A non-parametric test for linkage with a quantitative character. *Ann Hum Genet* 38:451–60.

[91] Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–16.

[92] Spiess EB (1977) *Genes in Populations.* John Wiley & Sons, New York.

[93] Stephens M, Donnelly KP (2000) Inference in molecular population genetics. *J R Stat Soc B* 62:605–655.

[94] Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73:1162–9.

[95] Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–89.

[96] Stram DO (2004) Tag SNP selection for association studies. *Genet Epidemiol* 27:365–374.

[97] Stumpf MPH (2004) Haplotype diversity and SNP frequency dependence in the description of genetic variation. *Eur J Hum Genet* 12:469–477.

[98] Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. *J Am Stat Assoc* 82:528–540.

[99] Tapper WJ, Maniatis N, Morton NE, Collins A (2003) A metric linkage disequilibrium map of a human chromosome. *Ann Hum Genet* 67:487–494.

[100] Teng J, Risch N (1999) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res* 9:234–41.

[101] Terwilliger JD, Hiekkalinna T (2006) An utter refutation of the "Fundamental Theorem of the HapMap". *Eur J Hum Genet* 14:426–437.

[102] Thomas DC, Stram DO (2006) An utter refutation of the 'Fundamental Theorem of the HapMap' by Terwilliger and Hiekkalinna. *Eur J Hum Genet* .

[103] Thompson CL, Rybicki BA, Iannuzzi MC, Elston RC, Iyengar SK, Gray-McGuire C, (SAGA) SGAC (2006) Reduction of sample heterogeneity through use of population substructure: an example from a population of African American families with sarcoidosis. *Am J Hum Genet* 79:606–613.

[104] Tiwari HK, Barnholtz-Sloan J, Wineinger N, Padilla MA, Vaughan LK, Allison DB (2008) Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Hum Hered* 66:67–86.

[105] Tu IP, Whittemore AS (1999) Power of association and linkage tests when the disease alleles are unobserved. *Am J Hum Genet* 64:641–649.

[106] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans C, Holt R, et al. (2001) The sequence of the human genome. *Science* 291:1304–51.

[107] Whittemore AS (2004) Estimating genetic association parameters from family data. *Biometrika* 91:219–225.

[108] Whittemore AS, Tu IP (2000) Detection of disease genes by use of family data. I. Likelihood-based theory. *Am J Hum Genet* 66:1328–40.

[109] Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet* 76:967–986.

[110] Yang BZ, Zhao H, Kranzler HR, Gelernter J (2005) Characterization of a likelihood based method and effects of markers informativeness in evaluation of admixture and population group assignment. *BMC Genet* 6:50.

[111] Zogel C, Boehringer S, Gross S, Varon R, Buiting K, Horsthemke B (2006) Identification of cis- and trans-acting factors possibly modifying the risk of epimutations on chromosome 15. *Eur J Hum Genet* 14:752–758.

[112] Zöllner S, Wen X, Hanchard NA, Herbert MA, Ober C, Pritchard JK (2004) Evidence for extensive transmission distortion in the human genome. *Am J Hum Genet* 74:62–72.