Human Heredity

Research Article

Hum Hered 2024;89:8–31
DOI: 10.1159/000535840

Received: September 11, 2023
Accepted: December 7, 2023
Published online: January 10, 2024

Joint Linkage and Association Analysis Using GENEHUNTER-MODSCORE with an Application to Familial Pancreatic Cancer

Markus Brugger[a, b, c] Manuel Lutz[a, b, c] Martina Müller-Nurasyid[a, b, c] Peter Lichtner[d] Emily P. Slater[e] Elvira Matthäi[e] Detlef K. Bartsch[e] Konstantin Strauch[a, b, c]

[a]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany; [b]Institute of Medical Information Processing, Biometry and Epidemiology - IBE, LMU Munich, Munich, Germany; [c]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [d]Institute of Human Genetics, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [e]Department of Visceral, Thoracic and Vascular Surgery, Philipps University, Marburg, Germany

Keywords
Association analysis · Familial pancreatic cancer · Haplotype frequency estimation · Linkage analysis · MOD scores

Abstract
Introduction: Joint linkage and association (JLA) analysis combines two disease gene mapping strategies: linkage information contained in families and association information contained in populations. Such a JLA analysis can increase mapping power, especially when the evidence for both linkage and association is low to moderate. Similarly, an association analysis based on haplotypes instead of single markers can increase mapping power when the association pattern is complex. ***Methods:*** In this paper, we present an extension to the GENEHUNTER-MODSCORE software package that enables a JLA analysis based on haplotypes and uses information from arbitrary pedigree types and unrelated individuals. Our new JLA method is an extension of the MOD score approach for linkage analysis, which allows the estimation of trait-model and linkage disequilibrium (LD) parameters, i.e., penetrance, disease-allele frequency, and haplotype frequencies. LD is modeled between alleles at a single diallelic disease locus and up to three diallelic test markers. Linkage information is contributed by additional multi-allelic flanking markers. We investigated the statistical properties of our JLA implementation using extensive simulations, and we compared our approach to another commonly used single-marker JLA test. To demonstrate the applicability of our new method in practice, we analyzed pedigree data from the German National Case Collection for Familial Pancreatic Cancer (FaPaCa). ***Results:*** Based on the simulated data, we demonstrated the validity of our JLA-MOD score analysis implementation and identified scenarios in which haplotype-based tests outperformed the single-marker test. The estimated trait-model and LD parameters were in good accordance with the simulated values. Our method outperformed another commonly used JLA single-marker test when the LD pattern was complex. The exploratory analysis of the FaPaCa families led to the identification of a promising genetic region on chromosome 22q13.33, which can serve as a starting point for future mutation analysis and molecular research in pancreatic cancer. ***Conclusion:*** Our newly proposed JLA-MOD score

karger@karger.com
www.karger.com/hhe

Correspondence to:
Markus Brugger, brugger@uni-mainz.de

method proves to be a valuable gene mapping and characterization tool, especially when either linkage or association information alone provide insufficient power to identify the disease-causing genetic variants.

Introduction

Traditionally, the identification of human disease genes is accomplished using the positional cloning approach, in which linkage analysis serves as the first step to narrow down the chromosomal position of the putative trait locus, followed by a fine-mapping association analysis [1]. Linkage analysis evaluates the co-segregation of genetic marker alleles together with a trait in families. Association analysis usually investigates the correlation of marker and disease-allele frequencies (linkage disequilibrium [LD]) between unrelated cases and controls on the population level (e.g., [2, 3]).

A joint linkage and association analysis (JLA) can substantially increase mapping accuracy and power because it makes use of both family and population information [4, 5]. In the following parts of the introduction, we give a brief review of linkage, association, and JLA methods. Subsequently, we introduce our newly proposed JLA method and describe the objective of the current paper.

Linkage Analysis

Linkage analysis has widely been used as the primary tool for the genetic mapping of traits with familial aggregation [6]. Methods of linkage analysis are commonly distinguished as either being parametric ("model-based") or nonparametric ("model-free"). In parametric linkage analysis, which is also known as model-based or LOD score analysis, a certain set of trait-model parameters is explicitly assumed for the segregation of the disease. Nonparametric linkage analysis methods proceed without explicit assumptions as to the trait-model parameters; however, it can be shown that certain nonparametric and parametric linkage tests are equivalent for a particular type of pedigree [7, 8]. In the simplest case of a diallelic autosomal trait locus causing a dichotomous disease, which is assumed throughout this paper, the trait-model parameters are the disease-allele frequency p_m ("m" for mutant, with wild-type allele frequency $p_+ = 1–p_m$) and the three penetrances f_0, f_1, and f_2, with f_i denoting the probability that an individual with i copies of the disease allele is affected by the disease. In addition, the recombination fraction θ between marker and trait locus, or the genetic position x of the putative trait locus in the case of a multipoint analysis, is modeled. The trait-model parameters can either be prespecified according to results from previous segregation analyses or maximized along with the recombination fraction in a joint segregation and linkage analysis. A so-called MOD score analysis allows researchers to jointly investigate segregation and linkage [9, 10] and avoids a potential loss in power due to model misspecifications that may occur in standard LOD score analysis [10]. Due to the maximization over trait-model parameters, MOD scores are inflated when compared to LOD scores. Since the asymptotic distribution of MOD scores is unknown in the general case, p values for the linkage test must be obtained by simulating the distribution of the MOD score under the null hypothesis of no linkage. Going beyond pure disease gene mapping, MOD score analysis can be used in gene characterization studies, which involve estimation of disease gene properties such as penetrance and disease-allele frequencies for ensuing risk calculations [11]. The core statistic of a MOD score analysis is the likelihood ratio of the pedigree likelihoods under the alternative hypothesis of linkage ($\theta \leq 0.5$) versus under the null hypothesis of no linkage ($\theta = 0.5$). The likelihood ratio is maximized with respect to θ as well as the trait-model parameters. It is of note that the same set of values for the trait-model parameters is used for the numerator as well as for the denominator of the likelihood ratio. As a consequence, the MOD score is proportional to the pedigree likelihood conditional on the trait phenotypes and hence leads to unbiased estimates of the trait-model parameters so that ascertainment through the trait is irrelevant [12]. However, this only holds for a linkage analysis in the absence of LD between marker and trait locus alleles and given a few other conditions summarized in Ginsburg et al. [13] and Malkin and Elston [14], which were reviewed and investigated for MOD score analysis in Brugger et al. [15]. The MOD score approach is implemented in the software package GENEHUNTER-MODSCORE (GHM) [16–19], which is maintained and continuously developed further by our working group. An implementation of the MOD score approach for quantitative trait loci, GENEHUNTER-QMOD, has been developed by Künzel and Strauch [20].

Association Analysis

Genetic association analysis tests for a correlation between disease status and genetic variation to identify putative disease genes [21]. Association analysis in pedigrees has traditionally been done using triads (case-parent trios) by comparing the probabilities of transmission for

each marker allele from the parents to their offspring under the assumption of complete linkage between marker and trait locus. The ascertainment of parents thereby enables a joint analysis of multiple marker loci with a more accurate assignment of the phase of the marker-locus alleles as compared to case-control data [22]. Such a procedure leads to a test for LD conditional on linkage, which has been formalized in the haplotype relative risk [23] and the haplotype-based haplotype relative risk method [24]. Moving from triads to larger sibships, the transmission/disequilibrium test TDT [25] and its extensions [26–35] are popular examples for nonparametric methods that draw information from both the linkage and association component. The original TDT approach [25] formally tests the null hypothesis of association but no linkage against the alternative of linkage in the presence of association in the analysis of multiple affected individuals from a single pedigree. When the analysis is restricted to independent triads, the null hypothesis of the TDT corresponds to no linkage or no association. Such methods, however, were originally designed for simple pedigree relationship structures and do not make use of any information regarding the mode of inheritance and trait-model parameters [36]. Several TDT-like approaches and extensions were implemented in software packages like FBAT [37, 38], PedGenie [39], QTDT [40], TRANSMIT [41], and UNPHASED [42]. Notably, Göring and Terwilliger [4] have shown how all abovementioned nonparametric association tests can be parametrized into a unifying likelihood framework, allowing for flexible likelihood ratio tests with different combinations for the null and alternative hypothesis.

Joint Linkage and Association (JLA) Analysis

A JLA analysis combines linkage and association information gathered from pedigrees, whereby association information on the population level can also be added using unrelated individuals. Linkage analysis methods generally make the assumption of linkage equilibrium (LE) between alleles at marker and disease loci. However, disease loci can be in LD with their flanking markers over a large distance, depending on their map distance and their population history [43]. Hence, the assumption of LE can reduce power of the linkage test when compared to a model that allows for LD [44]. On the other hand, if LD is present between alleles of the marker loci, assuming LE can increase the type I error of the linkage test in the case of missing parental genotypes [45–48]. Association analysis exploits LD information from the population; however, its power decays rapidly with increasing marker-trait locus distance, i.e., starting already from 1 centiMorgan [2]. Hence, it would be desirable to combine the two orthogonal mapping information components of linkage and association into a JLA analysis, which can have higher power compared to pure linkage or pure association analysis, especially when analyzing a dataset comprised of unrelated individuals and families [4, 5]. The idea of a JLA analysis is not new. Already in 1984, MacLean et al. [49] pointed out that such a JLA analysis can increase mapping power. In 1988, Clerget-Darpoux et al. [50] devised the MASC method, in which allelic association and segregation information is comprised in a χ^2 sum statistic. Later on, Tienari et al. [51] found that the incorporation of association into their LOD score linkage analysis dramatically increased power. Approaches of JLA analysis to map quantitative trait loci, which are not further considered in this work, can be found in Fan et al. [52] and Jung et al. [53].

In model-based analysis, incorporation of association information is achieved by including a parameter for LD between investigated genetic markers and the disease locus in the pedigree likelihood. Such methods, which can accommodate for association, have been implemented in popular software packages such as PAP [54] or jPAP [55] for segregation analysis and LINKAGE [56–58], MENDEL [59, 60], LAMP [61, 62], and PSEUDOMARKER [4, 63, 64] for linkage analysis. Although these implementations offer the ability to include association information into the calculations, formal joint tests for linkage and association are less common. A parametric, likelihood-based approach to JLA analysis was presented by Lou et al. [5, 65], who also pointed out that neglecting association information can lead to a loss in statistical power of the linkage test and to biased estimates of the recombination fraction. Another JLA approach, implemented in the PSEUDOMARKER software package, exploits the equivalence of parametric and nonparametric linkage methods and offers various likelihood ratio tests with different null and alternative hypotheses including a JLA test for single markers using twopoint calculations [4, 63, 64]. The JLA method of Xiong and Jin [36] is an extension to parametric LOD score analysis and has been implemented in MENDEL by Cantor et al. [66]. The likelihood-based framework implemented in the software package LAMP [61, 62] basically corresponds to a MOD score analysis (under some constraints) that includes association parameters and incorporates flanking marker information in a multipoint analysis. However, LAMP only performs likelihood ratio tests for pure linkage, for association conditional on linkage, and for the existence of further unobserved genetic variants apart from a trait

locus associated with the currently tested marker. In summary, an analysis that explicitly allows for a joint test of linkage and association using MOD scores is still lacking.

JLA Analysis Using MOD Scores

A MOD-score-based JLA analysis offers the joint estimation of the recombination fraction (or the genetic position in a multipoint setting), the penetrance function, and haplotype frequencies combining alleles of the disease locus and one or more marker loci. Although computationally demanding, such estimates can provide valuable insights into disease etiology and may contribute to improve genetic risk calculation and counselling [11]. In addition, the MOD score approach, as implemented in the GHM software package [67], accommodates for genomic imprinting – an epigenetic phenomenon that is known to play a role in a growing number of human diseases [68]. Imprinting is characterized by the dependence of an individual's liability to develop a disease according to the parental origin of the mutated allele(s). The ability of the MOD score approach to estimate trait-model parameters including the degree of imprinting depending on different pedigree types has been demonstrated in the context of linkage analysis [15, 69]. In the presence of LD, trait-model parameter estimates obtained from a MOD score analysis may be biased because sampling of pedigrees and individuals is no longer marker-independent, which is one of the necessary conditions of the ascertainment/sampling-assumption free property of the MOD score [12–14, 70], which are reviewed in [15]. However, the bias is argued to be only trivial [14, 70].

Linkage Information in JLA Analysis

Gathering linkage information from flanking markers in a multipoint calculation can increase mapping power in a JLA analysis as compared to a twopoint analysis [61]. However, usage of linkage information gathered from flanking markers has so far only been implemented in LAMP for LD tests conditional on linkage [61, 62].

Single-Marker versus Haplotype-Based Association Information in JLA Analysis

Another important aspect of JLA analysis is the question as to whether association information should be included from either a single marker or multi-marker haplotypes. There is evidence that haplotype-based association methods can outperform single-marker analysis [71], especially when there are multiple disease-causing alleles within the same gene and LD between the investigated markers is rather weak [72, 73]. However, haplotype-based methods are computationally expensive, especially in the case of missing genotypes, and result in a large number of additional degrees of freedom (*df*) for the likelihood ratio test, which might diminish power. Moreover, phase ambiguity of haplotypes needs to be handled by haplotype frequency estimation methods such as the expectation-maximization (EM) algorithm [74, 75] with the additional assumption of Hardy-Weinberg equilibrium in the population. Yet, the relative efficiency of single-marker versus haplotype-based approaches for modeling association is largely unexplored [73]. Remarkably, a JLA method to model LD between alleles at the trait locus and alleles at more than a single marker is implemented in MENDEL [66].

Objectives

The current work presents an extension of the MOD score approach which allows the joint analysis of linkage and association, using data from arbitrary pedigree types (extended pedigrees, nuclear families, triads, half-sibships) and unrelated individuals (singletons). We set out to implement this joint linkage and association extension (JLA-MOD score) in a new version of our GHM software package. To this end, LD was modeled by using one to three single nucleotide variants (SNVs) as test markers and by incorporating information for the linkage component from additional flanking markers with an arbitrary number of alleles.

In this paper, we thoroughly explain the details of the methodological advances and their implementation in the new GHM version 4. Then, we evaluate the type I error and power of the newly proposed JLA-MOD score using various simulation scenarios. In addition, we compare linkage and association parameter estimates obtained from the JLA-MOD score analysis with the simulated values. We also evaluate the relative mapping efficiency of new (JLA) and existing (pure linkage) GHM analysis options, depending on the underlying simulation scenario. In order to evaluate the costs and benefits of jointly estimating numerous linkage and LD parameters, we compare the type I error and power of the JLA-MOD score with the parsimonious JLA test implemented in the PSEUDOMARKER software [4, 63, 64]. The PSEUDOMARKER method proved to be a powerful approach in various types of linkage and/or association analyses, thereby outperforming many other methods [63, 64]. Lastly, we present a JLA-MOD score analysis using pedigree data from the German National Case Collection for Familial Pancreatic Cancer (FaPaCa) to demonstrate the applicability of our new method in practice.

Methods

Extension of the MOD Score Likelihood Ratio to Accommodate for LD

In pure linkage analysis assuming a dichotomous trait, which is governed by a diallelic locus, the MOD score is defined as the ratio of the likelihoods under the alternative hypothesis of linkage and the likelihood under the null hypothesis of no linkage, maximized over the trait-model parameters (penetrances f_0, f_1, f_2 and disease-allele frequency p_m) and the recombination fraction θ (or, in the case of a multipoint analysis, the genetic position x):

$$\mathrm{MOD} = \max_{p_m, f_0, f_1, f_2, \theta} \log_{10} \frac{L(p_m, f_0, f_1, f_2, \theta)}{L(p_m, f_0, f_1, f_2, \theta = 0.5)} \quad (1)$$

As mentioned in the Introduction section, the same set of values for the trait-model parameters is used for the numerator as well as for the denominator of the likelihood ratio. If imprinting is modeled, f_1 is split up into two heterozygote penetrances, $f_{1,\,\mathrm{pat}}$ and $f_{1,\,\mathrm{mat}}$, according to the origin of the parental allele [67]. In order to accommodate for association information, the likelihood is extended to include a parameter for LD:

$$\mathrm{MOD} = \max_{p_m, f_0, f_1, f_2, \theta, LD} \log_{10} \frac{L(p_m, f_0, f_1, f_2, \theta, LD)}{L(p_m, f_0, f_1, f_2, \theta = 0.5, LD = 0)} \quad (2)$$

It is of note that the recombination fraction θ is confounded with the allele sharing at the marker locus and hence also with the trait-model parameters [76], which is commonly avoided by assuming no recombination between marker and trait locus [61]. Maximization over θ, or the genetic position x, is nevertheless performed in practice by evaluating (1) or (2) for different genetic positions. Linkage information is represented by the distribution of inheritance vectors, which represent the patterns of founder allele segregation in a pedigree, for a given genetic position. The inheritance vector contains 1 bit for each meiosis in the pedigree, with 0 and 1 denoting transmission of the paternally or maternally inherited allele, respectively. The distribution of inheritance vectors can be obtained using a hidden Markov model in the context of the Lander-Green algorithm [77], which is used in GHM. The Lander-Green algorithm scales linearly with the number of analyzed markers but is limited to the analysis of modestly sized pedigrees. Brief reviews of the Lander-Green algorithm are given in [19, 78]. The distribution of all inheritance vectors is calculated assuming a particular position of the trait locus relative to a marker or group of markers. In the case of no linkage, the distribution is uniform, whereas under linkage, it is usually peaked at few inheritance vectors that are compatible with the observed marker alleles. This distribution under the assumption of linkage contributes to the numerator of (1) and (2), whereas the case of no linkage ($\theta = 0.5$) with a uniform inheritance-vector distribution contributes to the denominator of (1) and (2).

Parametrization of LD

In the case of a single test SNV and a diallelic trait locus (TL), there are $2 \times 2 = 4$ haplotypes for all combinations of marker-trait locus alleles, namely: SNV|TL $\in$ {0|0, 0|1, 1|0, 1|1} = : $\{h_0, \ldots, h_3\}$, whereby 0 and 1 represent allele codes for the SNV and the trait locus alleles, with the wild-type allele "+" coded as 0 and with the mutant allele "m" coded as 1. LD can be parametrized by the respective haplotype frequencies $p_{h_0}, \ldots, p_{h_3}$ in the numerator of equation (2). The denominator of (2) models LE, i.e., independence of marker and trait locus alleles, by separate contributions of the test SNV haplotype frequencies (or allele frequencies in the case of a single test SNV) and of the disease (or wild-type) allele frequency to the likelihood. In pedigree and/or singleton likelihood analysis, it is advisable to estimate marker-haplotype frequencies directly from the data under study [23, 79, 80], which can be achieved using the EM algorithm (see [78]). The obtained values serve as marker-haplotype frequencies (or allele frequencies for a single test SNV) in the denominator of equation (2). This way, allele or haplotype frequencies for the marker data are estimated before maximizing equation (2), leaving the disease out of the analysis in the first place. This yields estimates that are identical to those obtained in a joint analysis of trait and marker phenotypes when there is in fact no linkage [80]. In the case of a single test SNV, the EM-estimated allele frequencies are denoted by $p_{h_0}^{EM}$ and $p_{h_1}^{EM}$ for SNV alleles 0 and 1, respectively, whereby $p_{h_0}^{EM} + p_{h_1}^{EM} = 1$. Plugging all these frequencies in equation (2), the MOD score then reads:

$$\mathrm{MOD} = \max_{p_m, f_0, f_1, f_2, \theta, p_{h_1}, p_{h_2}, p_{h_3}} \log_{10} \frac{L(p_m, f_0, f_1, f_2, \theta, p_{h_1}, p_{h_2}, p_{h_3})}{L\left(p_m, f_0, f_1, f_2, \theta = 0.5, p_{h_0}^{EM}, p_{h_1}^{EM}\right)} \quad (3)$$

Here, p_+ and p_{h_0} can be omitted from the formula due to the restrictions $p_m + p_+ = 1$ and $\sum_{i=0,..,3} p_{h_i} = 1$. Further, $\sum_{i=0,2} p_{h_i} = p_+$, and $\sum_{i=1,3} p_{h_i} = p_m$. Note that the SNV frequencies $p_{h_0}^{EM}$ and $p_{h_1}^{EM}$ do not correspond to the marginal allele frequencies that can be calculated from the numerator frequencies $p_{h_1}, p_{h_2}, p_{h_3}$ and p_{h_0}, but instead are fixed values during the maximization of equation (3) (see also above).

With two test SNVs, there are eight marker-trait haplotypes: $\mathrm{SNV}_1|\mathrm{SNV}_2|\mathrm{TL} \in \{0|0|0, 0|0|1, 0|1|0, 0|1|1, 1|0|0, 1|0|1, 1|1|0, 1|1|1\} = : \{h_0, \ldots, h_7\}$. The respective haplotype frequencies are denoted by $p_{h_0}, \ldots, p_{h_7}$. The corresponding EM-estimated marker-haplotype frequencies are given by $p_{h_0}^{EM}, \ldots, p_{h_3}^{EM}$. The MOD score then reads:

$$\mathrm{MOD} = \max_{p_m, f_0, f_1, f_2, \theta, p_{h_1}, \ldots, p_{h_7}} \log_{10} \frac{L(p_m, f_0, f_1, f_2, \theta, p_{h_1}, \ldots, p_{h_7})}{L\left(p_m, f_0, f_1, f_2, \theta = 0.5, p_{h_0}^{EM}, \ldots, p_{h_3}^{EM}\right)} \quad (4)$$

Here, p_+ and p_{h_0} can again be omitted from the formula due to the restrictions $p_m + p_+ = 1$ and $\sum_{i=0,\ldots,7} p_{h_i} = 1$. Further, $\sum_{i=0,2,4,6} p_{h_i} = p_+$, and $\sum_{i=1,3,5,7} p_{h_i} = p_m$. The EM-estimated marker-haplotype frequencies in the denominator of equation (4) are again fixed values and are constant during the maximization of the likelihood ratio.

In the case of three test SNVs, there are 16 marker-trait locus haplotypes: $\mathrm{SNV}_1|\mathrm{SNV}_2|\mathrm{SNV}_3|\mathrm{TL} \in \{0|0|0|0, 0|0|0|1, 0|0|1|0, 0|0|1|1, 0|1|0|0, 0|1|0|1, 0|1|1|0, 0|1|1|1, 1|0|0|0, 1|0|0|1, 1|0|1|0, 1|0|1|1, 1|1|0|0, 1|1|0|1, 1|1|1|0, 1|1|1|1\} = : \{h_0, \ldots, h_{15}\}$. The respective haplotype frequencies are denoted by $p_{h_0}, \ldots, p_{h_{15}}$. The EM-estimated marker-haplotype frequencies are given by $p_{h_0}^{EM}, \ldots, p_{h_7}^{EM}$.

The MOD score for three test SNVs then reads:

$$\mathrm{MOD} = \max_{p_m, f_0, f_1, f_2, \theta, p_{h_1}, \ldots, p_{h_{15}}} \log_{10} \frac{L(p_m, f_0, f_1, f_2, \theta, p_{h_1}, \ldots, p_{h_{15}})}{L(p_m, f_0, f_1, f_2, \theta = 0.5, p_{h_0}^{EM}, \ldots, p_{h_7}^{EM})} \quad (5)$$

Here, p_+ and p_{h_0} can again be omitted from the formula due to the restrictions $p_m + p_+ = 1$ and $\sum_{i=0,\ldots,15} p_{h_i} = 1$. Further, $\sum_{i=0,2,4,6,8,10,12,14} p_{h_i} = p_+$, and $\sum_{i=1,3,5,7,9,11,13,15} p_{h_i} = p_m$. The EM-estimated marker-haplotype frequencies in the denominator of equation (5) are again fixed values and are constant during the maximization of the likelihood ratio. More detailed constraints for the linkage and LD parameters are provided below. It is of note that singletons and triads only contribute association information in terms of haplotype frequencies to the likelihood, whereas pedigrees contribute both linkage and association information. The MOD score for the complete dataset is obtained by summing the log-likelihood ratios in equation (3), (4), or (5) over all pedigrees and singletons in the dataset, with the maximization being performed over the sum.

Detailed Formulation of the MOD Score Likelihood Ratio

The likelihood ratios for each pedigree in equations (3), (4), and (5) can be rewritten in terms of scoring functions for the inheritance vectors v at a given genetic position, as well as the inheritance-vector distributions under linkage and no linkage:

$$\mathrm{MOD} = \log_{10} \frac{\sum_v Scoring_1(v) \cdot P_{complete}(v)}{\left(\sum_v Scoring_2(v) \cdot P_{complete}(v)\right) \cdot \left(\sum_v Scoring_3(v) \cdot P_{uniform}(v)\right)} \quad (6)$$

Without loss of generality, the following details are explained for the case of a single test SNV:

- $Scoring_1(v)$ contains the product over penetrances for all $f+n$ individuals in a pedigree (with f denoting the number of founders and n denoting the number of nonfounders) and marker-trait locus haplotype frequencies $p_{h_0}, \ldots, p_{h_3}$ for all f founders in a pedigree, given a set of ordered founder genotypes (*OFG*) of the test SNV and the disease locus as well as ordered nonfounder genotypes (*ONG*) as assigned by the *OFG*s together with the inheritance vector v. The sum is then taken over those of the $2^{2f} \times 2^{2f}$ possible *OFG*s that are compatible with the observed test SNV genotypes of all individuals in the pedigree:

$$Scoring_1(v) = \sum_{\substack{OFG \\ compatible}} \prod_{k \in \mathscr{F}} p_{h_{OFG_{k,1}}} p_{h_{OFG_{k,2}}} f_{g(OFG_k)} \prod_{k \in \mathscr{N}} f_{g(ONG_k(OFG,v))}$$

$\mathscr{F}$ represents the set of founders and $\mathscr{N}$ the set of nonfounders in the pedigree. $p_{h_{OFG_{k,1}}}$ and $p_{h_{OFG_{k,2}}}$ are the marker-trait locus haplotype frequencies for founder individual k of the paternally and maternally inherited haplotypes, respectively, with $OFG_{k,1}, OFG_{k,2} \in \{0,1,2,3\}$. $f_{g(OFG_k)}$ denotes the penetrance of founder individual k according to the disease genotype $g \in \{0, \text{“1, pat”}, \text{“1, mat”}, 2\}$, which is a function of the ordered genotype OFG_k (comprising test SNV and disease locus) of founder individual k. $f_{g(ONG_k(OFG,v))}$ denotes the penetrance of nonfounder individual k according to the disease genotype g, which is a function of the ordered genotype ONG_k (comprising test SNV and disease locus) of nonfounder individual k, which again depends on the given set of ordered founder genotypes (*OFG*) together with the inheritance vector v. In the case of genomic imprinting, the ordered genotype formulation allows us to define different penetrances for individuals heterozygous at the disease locus by taking the parental origin of the mutant allele into account. The ordered founder genotypes are directly assigned within the summation, and the ordered nonfounder genotypes are determined by the ordered founder genotypes together with the inheritance vector.

The algorithm to filter out ordered founder genotypes that are compatible with the observed SNV genotypes of all individuals in a pedigree and the inheritance vector is explained in the context of the haplotype frequency estimation in the next section.

- $P_{complete}(v)$ denotes the probability for an inheritance vector v based on the inheritance distribution at a given genetic position conditional on the additional flanking markers, i.e., the markers beyond the one, two, or three SNVs tested for LD with the putative disease locus, as obtained by the Lander-Green algorithm.
- $Scoring_2(v)$ denotes the product over the allele frequencies of the test SNV, or haplotype frequencies in the case of two or three test SNVs, for all f founders in a pedigree:

$$Scoring_2(v) = \sum_{\substack{OFSG \\ compatible}} \prod_{k \in \mathscr{F}} p_{h_{OFSG_{k,1}}}^{EM} \, p_{h_{OFSG_{k,2}}}^{EM}$$

where *OFSG* denotes a particular set of ordered test SNV genotypes for all founders, $p_{h_{OFSG_{k,1}}}^{EM}$ and $p_{h_{OFSG_{k,2}}}^{EM}$ are the test SNV allele frequencies for founder individual k of the paternally and maternally inherited alleles, respectively, with $OFSG_{k,1}, OFSG_{k,2} \in \{0,1\}$, and the sum is taken over all sets of ordered test SNV genotypes that are compatible with the observed genotypes.

- $Scoring_3(v)$ denotes the product over penetrances for all $f+n$ individuals in a pedigree and disease-allele frequencies for all f founders given a set of ordered founder disease genotypes (*OFDG*). The sum is then taken over all 2^{2f} possible *OFDG*s:

$$Scoring_3(v) = \sum_{OFDG_{k \in \mathscr{F}}} \prod p_{OFDG_{k,1}} p_{OFDG_{k,2}} f_{g(OFDG_k)} \prod_{k \in \mathscr{N}} f_{g(ONDG_k(OFDG,v))}$$

with $p_{OFDG_{k,1}}$ and $p_{OFDG_{k,2}}$ and denoting the disease-locus allele frequencies for founder individual k of the paternally and maternally inherited alleles, respectively, with $OFDG_{k,1}, OFDG_{k,2} \in \{+, m\}$. $f_{g(OFDG_k)}$ denotes the penetrance of founder individual k according to the disease genotype $g \in \{0, \text{“1, pat”}, \text{“1, mat”}, 2\}$, which is a function of the ordered disease genotype $OFDG_k$ of founder individual k. $f_{g(ONDG_k(OFDG,v))}$ denotes the penetrance of nonfounder individual k according to the disease genotype $g \in \{0, \text{“1, pat”}, \text{“1, mat”}, 2\}$, which is a function of the ordered disease genotype $ONDG_k$ of nonfounder individual k, which depends on the given set of ordered founder disease genotypes (*OFDG*) together with the inheritance vector v.

- $P_{\mathrm{uniform}}(v)$ denotes the probability for inheritance vector v based on the inheritance distribution at a given genetic position of the putative disease locus under no linkage with the markers. The inheritance distribution under the null hypothesis of no linkage is uniform, i.e., all inheritance vectors are equally likely.

Combining $Scoring_3(v)$ with $P_{\mathrm{uniform}}(v)$ reflects the fact that the trait locus is unlinked to the underlying genetic position and the marker locus. Conversely, the test SNV remains at its original genetic position, which is reflected by combining $Scoring_2(v)$ with

$P_{complete}(v)$. In summary, identical to equations (3), (4), and (5), the numerator of equation (6) reflects the alternative hypothesis of linkage and association of the disease locus with the markers. The denominator reflects the null hypothesis of no linkage and no association, for which the disease locus is assumed to be at a position resulting in complete independence with regard to allelic correlation and co-segregation.

Haplotype Frequency Estimation

In GHM 4, marker-allele and marker-haplotype frequencies are directly estimated from the data under study using a gene-counting based EM algorithm. To this end, haplotype frequencies for clusters of up to three tightly linked SNVs in a given test set as well as allele frequencies for flanking markers with two or more alleles can be estimated. The recombination fraction between test SNVs of a given cluster is assumed to be 0, SNVs within a cluster can exhibit any degree of LD, and missing genotypes are allowed for founders and nonfounders. Standard algorithms for the estimation of haplotype frequencies for independent observations of a population can readily be extended to include pedigree information, which improves haplotype frequency estimates for the general population by exclusion of nonexistent haplotype configurations from the analysis [81]. The haplotype frequency estimation in pedigrees is applied over the independent parents, whereby their children's genetic phenotypes are used to exclude those haplotype pairs from the analysis, which are possible for the founders, but contradictory for the children [81]. An implementation of such a procedure in the context of the Lander-Green algorithm to compute the haplotype-based disease-locus likelihood in pure linkage analysis was presented by Abecasis and Wigginton [78] for the linkage analysis software package Merlin [82]. As GHM is also based on the Lander-Green algorithm, our implementation of the haplotype frequency estimation is similar to the method described in [78]. Noteworthy, the original GENEHUNTER software also offers methods to identify the most likely haplotypes for each pedigree using the Lander-Green and the Viterbi algorithm [83]; since GHM is based on GENEHUNTER, these haplotyping methods have been available in former versions of GHM as well. A general overview of haplotyping methods for pedigrees can be found in [84].

The first step of our newly implemented haplotype frequency estimation algorithm corresponds to the enumeration of the entire set of inheritance vectors. Since there are $2n$ meioses in a pedigree, with n denoting the number of nonfounders, there are 2^{2n} inheritance vectors [77], which can be reduced to 2^{2n-f} identifiable inheritance vectors for the analysis, with f denoting the number of founders in a pedigree [83]. Second, the algorithm iterates over all inheritance vectors and markers of the SNV test set to calculate the probability of the observed genotypes for each marker conditional on a particular inheritance vector, which essentially reduces to a product of haplotype frequencies with two frequencies for each founder in the pedigree. This step is achieved by identifying all ordered founder genotypes that are compatible with the observed founder genotypes of a given marker. Next, the conditional probability of the genotypes of all individuals in the pedigree given an ordered, and hence phased, founder genotype configuration, i.e., of founder haplotypes, and a given inheritance vector is calculated for a given marker of the test set by genetic descent-graph analysis [85]. Briefly, phased founder alleles are assigned to all offspring in the pedigree using the inheritance vector. The correspondingly assigned nonfounder genotypes are compared to the observed genotypes. The conditional probability of the genotypes, given a phased founder genotype configuration, then simply takes on the value 1 for a compatible or 0 for an incompatible genotype. These steps are repeated for all markers of a given set of test SNVs. Finally, the Cartesian product of all identified possible phased founder genotypes for a given inheritance vector across all markers of the test set leads to the set of compatible founder haplotype configurations for this particular inheritance vector. This process of reducing the space of possible founder haplotype configurations by descent-graph analysis is also called diplotype reduction [86], for which an illustrative example in the context of the Lander-Green algorithm can be found in [78]. If the set of noncontradictory haplotype configurations for a given pedigree is empty, there either is an error in the genotypes or relationships in the pedigree, or a recombination event happened. Although a recombination event can contain valuable information [81], the haplotype frequency calculation cannot proceed in this case. However, with closely linked SNVs and modestly sized pedigrees, recombination events should be rare, even at higher recombination fractions [81]. The aforementioned steps are repeated for all s pedigrees in the sample. During the generation of the set of noncontradictory haplotype configurations, different inheritance vectors will likely yield the same configurations, such that calculations can be saved by incrementing a coefficient for the number of appearances of a particular configuration for different inheritance vectors [78]. The results of these calculations are generic, i.e., not specific for a particular set of haplotype frequencies and are then used in the following EM algorithm, which involves two basic steps. First, the expected number of haplotype copies is estimated, conditional on current haplotype frequency estimates. Next, these expected counts are used to obtain new haplotype frequencies. Repeatedly updating haplotype frequencies and estimated counts in turn finally converges to maximum-likelihood estimates for the haplotype frequencies. Convergence to local optima can be controlled by assuming different sets of starting values for the first EM iteration. In GHM 4, two sets of initial values for the haplotype frequencies are applied to monitor convergence. In the case of a single test SNV, the EM algorithm is initialized in a first run with equal allele frequencies and in a second run with the frequencies provided in the marker data file. In the case of two and three test SNVs, the EM algorithm is initialized in a first run with equal haplotype frequencies and in a second run with the product of single-marker-allele frequencies, which were estimated beforehand using a separate round of the EM algorithm. Given a set of initial values for the haplotype frequencies p_{h_r}, $r = 0, \ldots, 2^m-1$ for m SNVs in the test set, F founders in all s pedigrees, with f founders in each pedigree, the recursion formula of the EM algorithm for frequency p_{h_r} at iteration $t+1$ is:

$$p_{h_r}^{EM(t+1)} = \frac{1}{2F} \sum_{s} \frac{\sum_{\substack{OFSG \\ compatible}} z_{OFSG}^{h_r} c_{OFSG} \prod_{k \in \mathcal{F}} p_{h_{OFSG_{k,1}}}^{EM(t)} \; p_{h_{OFSG_{k,2}}}^{EM(t)}}{\sum_{\substack{OFSG \\ compatible}} c_{OFSG} \prod_{k \in \mathcal{F}} p_{h_{OFSG_{k,1}}}^{EM(t)} \; p_{h_{OFSG_{k,2}}}^{EM(t)}} \qquad (7)$$

where *OFSG* denotes a particular set of ordered test SNV genotypes for all founders, and $p^{EM(t)}_{h_{OFSG_{k,1}}}$ and $p^{EM(t)}_{h_{OFSG_{k,2}}}$ are the haplotype frequencies for founder individual k of the paternally and maternally inherited haplotypes at the previous iteration t, respectively, with $OFSG_{k,1}, OFSG_{k,2} \in \{0, \ldots, 2^m - 1\}$. $z^{h_r}_{OFSG}$ counts the number of appearances of haplotype h_r in the given *OFSG*, c_{OFSG} is the coefficient counting the number of different inheritance vectors compatible with *OFSG*, and $\mathscr{F}$ represents the set of founders in a single pedigree. The iteration stops as soon as the haplotype frequencies, or equivalently the log-likelihood function, do not further improve by a predefined accuracy limit. The log-likelihood function of the marker data is necessary to compare different EM solutions obtained using different initial values. The corresponding marker log-likelihood for equation (7) is given by:

$$\log(L_{marker}) = \sum_s \log\left(\frac{\sum_{\substack{OFSG \\ compatible}} c_{OFSG} \prod_{k \in \mathscr{F}} p^{EM}_{h_{OFSG_{k,1}}} p^{EM}_{h_{OFSG_{k,2}}}}{2^{2n-f}} \right)$$

Parameter Constraints for the MOD Score Calculation

In accordance with former GHM versions, the user can specify the disease-allele frequency to be bound within a certain range, typically not larger than 0.5 (default value). With regard to the penetrances, the user can set the restriction $f_0 \leq f_1 \leq f_2$ (default setting). The user can also allow for imprinting models, for which $f_{1,pat} \neq f_{1,mat}$ (default: $f_{1,pat} = f_{1,mat}$, i.e., no imprinting). With regard to the marker-trait locus haplotype frequencies, the constraints are coupled to the constraint imposed on the disease-allele frequency. Without any prespecified restriction, the general constraints are

$$p_m \in [0, 1]$$
$$p_{h_i} \in [0, 1]$$
$$\sum_{i=0,2,\ldots} p_{h_i} = p_+$$
$$\sum_{i=1,3,\ldots} p_{h_i} = p_m$$
$$\sum_{i=0,2,\ldots} p_{h_i} + \sum_{i=1,3,\ldots} p_{h_i} = p_+ + p_m = 1$$

with $\sum_{i=0,2,\ldots} p_{h_i}$ corresponding to the sum of those marker-trait locus haplotype frequencies p_{h_i} that carry the wild-type allele of the trait locus with marginal frequency p_+, and with $\sum_{i=1,3,\ldots} p_{h_i}$ corresponding to the sum of those marker-trait locus haplotypes that carry the mutant disease allele of the trait locus with marginal frequency p_m. The marker-locus haplotype frequencies in the denominator of the MOD score are obtained from the previous maximum-likelihood estimation and remain fixed in the denominator during the maximization of the likelihood ratio (see also above).

Maximization Routine for the JLA-MOD Score

GHM 4 maximizes the likelihood ratio using a two-step approach. First, a predefined grid of values for the disease-allele frequency and the penetrances is applied. The parameter set, containing a particular combination of the disease-allele frequency and the penetrances, is complemented with values for the p_{h_i} randomly drawn, such that all abovementioned parameter constraints are satisfied.

The initial grid-based MOD score, which is obtained by taking the highest score over all parameter sets, serves as the starting point for the second step of the maximization routine of GHM 4. In this second step, GHM 4 uses the local derivative-free, direct-search optimization method COBYLA ("Constrained Optimization BY Linear Approximations") that models the objective as well as any linear and non-linear equality and inequality constraint functions by linear interpolations [87, 88]. GHM 4 uses the COBYLA implementation in the programming language C, which is part of the free/open-source library NLopt ("Non-Linear Optimization") (v2.6.2) [89]. The algorithm operates by evaluating the objective function and the constraints at the vertices of a trust region. If the optimization problem has a total of N parameters, then the trust region has a total of $N+1$ vertices [90]. With this information, linear approximations of the objective function and constraints are employed during the optimization process. The strength of COBYLA lies in its robustness, which makes it a suitable tool for noisy functions [90]. In GHM 4, COBYLA is initialized by the set of parameters that led to the highest score of the grid-based maximization, and the return value represents the final MOD score. To improve convergence, the otherwise deterministic COBYLA algorithm is initialized with different initial step sizes for the parameters.

Moreover, the user can also specify fixed sets of trait-model parameters (disease-allele frequency and penetrances), for which individual MOD scores are calculated. In this case, the maximization routine works as described above, but optimizes only the marker-trait locus haplotype frequencies.

Construction of Test Marker Sets

The general assumption of LE between flanking markers in the calculations (i.e., between markers beyond the test SNVs) stays untouched in GHM 4. Sorting out flanking markers that are in LD with each other, which is most common when using dense SNVs, should be done prior to the analysis using selection methods as described in [91]. Diallelic SNVs can be used either as test SNVs or as flanking markers, the latter contributing linkage information only. Accordingly, two additional input files need to be specified for a JLA analysis: one containing a list of markers used for the multipoint linkage calculation ("flanking markers") and one containing a list of association regions, defined by the two outermost SNVs, for which all combinations of SNVs ("test markers") within a user-specified genetic distance are considered for building haplotypes of a given size (one, two, or three test SNVs per haplotype). The assignment of a SNV to both flanking and test marker sets is automatically recognized and ruled out. In the case of a recombination event, the current test set will be discarded with a suggestion to the user to reduce the maximum genetic distance between test SNVs. Alternatively, the user may specify a fixed test marker set of a particular size (one, two, or three test SNVs) for JLA analysis, which can also be combined with specifying fixed sets of trait-model parameters.

Simulation of p *Values*

Because the distribution of JLA-MOD scores under the null hypothesis of no linkage and no association is unknown, p values for statistical inference must be obtained by simulations. To this end, GHM 4 offers an option to calculate a point-wise p value for

the JLA test using a particular set of test SNVs, which may have been identified during a previous JLA analysis with potentially many sets of test SNVs. The simulation run can be started using the same input files as for the initial JLA analysis, except that the user needs to specify the number of replicates and the test marker set of interest in a slightly adapted GHM commands file. GHM 4 offers parallel analysis of replicates, so that the user can specify the number of parallel processes as required for the simulation. Replicates can be stored on demand or reproduced by specifying the same random seed. The simulation algorithm works as follows. First, flanking marker and test marker genotypes are drawn for the founders based on the corresponding frequency distributions, which were estimated using the EM algorithm. Flanking marker and test marker genotypes are assigned to the offspring by gene-dropping, i.e., independent of disease status, according to the underlying genetic map. Ungenotyped individuals stay ungenotyped. The p value for the real dataset is calculated according to $p = \frac{k+1}{n+1}$, with n being the total number of replicates and k the number of replicates showing a MOD score that is equal to or higher than the one obtained from the real dataset.

Data Simulation and Analysis

Simulation Scenarios

In order to evaluate the new JLA analysis option in GHM 4, we simulated datasets consisting of small to moderately sized pedigrees and unrelated individuals. Specifically, 20 affected sib-pairs, 20 discordant sib-pairs (a sib-pair consisting of an affected and an unaffected sibling), 40 affected half-sib pairs (20 with a common mother, 20 with a common father), two three-generation pedigrees (3-Gs), 20 triads, 20 affected unrelated individuals (cases), and 20 non-affected unrelated individuals (controls) were simulated. Two trait models were considered. Trait model 1 (TM1) was simulated using a disease-allele frequency $p_m = 0.01$ and penetrances $f_0 = 0.01$; $f_1 = 0.09$; $f_2 = 0.17$. In addition, a second trait model (TM2) with maternal imprinting was simulated, also using a disease-allele frequency of 0.01, with penetrances according to the parental origin of the disease allele: $f_0 = 0.01$; $f_{1,\text{pat}} = 0.14$; $f_{1,\text{mat}} = 0.04$; $f_2 = 0.17$. With respect to the test markers, we simulated three perfectly linked SNVs with minor allele frequencies (MAFs) set to 0.1 for all three SNVs. Pairwise LD between alleles at the test markers was set to $D' = 0.5$. LD as measured by Cramér's V (see, e.g., [92]) between the three-SNV marker haplotypes and alleles at the diallelic trait locus was set to 0 for the simulations under the null hypothesis of no linkage and no association ($H_{0,\,a}$, with $\theta = 0.5$ between SNVs and trait locus) and also under the null hypothesis of linkage, but no association ($H_{0,\,b}$, with $\theta = 0$ between SNVs and trait locus). Hence, the corresponding values of Cramér's V between either the single-marker alleles or the 2-marker SNV haplotypes, for which either one or two SNVs were selected out of the three SNVs, and the alleles at the disease locus were also 0. Under the alternative hypothesis of linkage and association (H_1, with $\theta = 0$ between SNVs and trait locus), three patterns of LD were considered to investigate the statistical efficiency of modeling LD with 2- or 3-marker haplotypes, as compared to single-marker JLA or pure linkage analyses. Scenario S1 was designed as an example in which a single-marker analysis is sufficient to capture the LD pattern, resulting in no further advantage of the 2- and 3-marker haplotype analyses. Cramér's V was set to 0.158 between alleles of a single SNV and alleles at the trait locus. The corresponding Vs for the 2- and 3-marker haplotype formulations were 0.158 and 0.16, respectively. Scenario S2 was designed as an example in which the LD pattern is best captured by a 2-marker analysis, rendering it superior over the single- and 3-marker haplotype analyses. Cramér's V was set to 0.175 between haplotypes of two SNVs and alleles at the trait locus. The corresponding Vs for the single- and 3-marker haplotype formulations were 0.118 and 0.187, respectively. Finally, scenario S3 was designed as an example in which the 3-marker analysis is needed to fully capture the LD pattern, resulting in an advantage over the single- and 2-marker haplotype analyses. Cramér's V was set to 0.474 between haplotypes of three SNVs and alleles at the trait locus. The corresponding Vs for the single- and 2-marker haplotype formulations were 0.141 and 0.201, respectively.

As to the flanking markers, ten SNVs with a MAF of 0.1 were simulated in LE with each other on either side of the trait locus with $\theta = 0.002$ between each other and with $\theta = 0.001$ between the innermost flanking marker on each side and the trait locus, for both trait models and all LD scenarios. An overview of the simulated scenarios is given in Table 1. The population haplotype frequencies of the SNVs used for the simulation of marker data in the three LD scenarios can be found in Tables 2 and 3.

Simulation of Genotype Data

Generation of genotype data with or without imprinting effects and conditional on affection status was either carried out using SLINK [93–95] or by its imprinting extension SLINK Imprinting [96]. The simulation algorithm calculates the probability distribution of genotypes $\boldsymbol{g} = g_1, g_2, \ldots, g_n$ conditional on the phenotype values $\boldsymbol{x} = x_1, x_2, \ldots, x_n$ of n family members in a step-wise manner until all members have been assigned a genotype, each conditional on all phenotypes and the set of genotypes assigned before to other family members: $P(\boldsymbol{g}|\boldsymbol{x}) = P(g_1|\boldsymbol{x})P(g_2|g_1, \boldsymbol{x})P(g_3|g_1, g_2, \boldsymbol{x})\ldots$ The calculation time of this algorithm increases linearly with additional family members, but exponentially with the number of markers. In order to speed-up multi-marker simulations, a two-step algorithm originally developed by Lemire [97] was employed, which exploits the ability of conditional simulations by SLINK and SLINK Imprinting and uses a gene-dropping algorithm implemented in the SLINK utility program SUP [95, 97] to quickly generate a large number of markers. The first step of the algorithm generates disease-locus genotypes and trait values using SLINK or SLINK Imprinting. In the second step, SUP simulates flanking and test marker genotypes, taking into account the scenario-specific LD pattern between alleles at the test marker and trait loci.

Assessing Statistical Significance in JLA Analysis

For each scenario in Table 1, 1,000 datasets were simulated as described in the preceding section. p values were obtained using 999 replicates for each of the 1,000 datasets by applying the new simulation routine of GHM 4.

Investigated Test Approaches

In order to assess the statistical efficiency of our newly developed haplotype analysis approach, all scenarios were analyzed using pure linkage MOD score analysis with the previous GHM version 3 (GHM-MOD) and the newly proposed GHM-JLA analysis (GHM-JLA) using either one, two, or three test SNVs for the construction of test marker haplotypes. The same datasets simulated with three test SNVs were used as the basis for all three LD scenarios. In the case of the pure linkage and single-marker JLA analysis, the analysis was performed using the central test SNV only. In the case of the 2-marker

Table 1. Overview of the simulated scenarios to evaluate the statistical properties of the JLA-MOD score

Trait models and SNV scenarios								
TM1 $\theta \in \{0.0;0.5\}$; $p_m = 0.01$; $f_0 = 0.01$; $f_{1,pat} = 0.09$; $f_{1,mat} = 0.09$; $f_2 = 0.17$ Dominance index $D = 0$; Imprinting index $I = 0$								
TM2 $\theta \in \{0.0;0.5\}$; $p_m = 0.01$; $f_0 = 0.01$; $f_{1,pat} = 0.14$; $f_{1,mat} = 0.04$; $f_2 = 0.17$ Dominance index $D = 0$; Imprinting index $I = 0.625$								
3 test SNVs with $\theta = 0.0$ between SNVs								
Test SNVs	MAF_1	MAF_2	MAF_3	SNV-SNV LD (D')				
	0.1	0.1	0.1	0.5				
					LD (Cramér's*V*)			
					$H_{0,a}$	$H_{0,b}$	H_1	
							S1	S2	S3
				1-SNV-trait-locus LD	0.0	0.0	0.158	0.118	0.141
				2-SNVs-trait-locus LD	0.0	0.0	0.158	0.175	0.201
				3-SNVs-trait-locus LD	0.0	0.0	0.160	0.187	0.474
10 flanking SNVs on either side of the test SNVs with $\theta = 0.002$ between flanking SNVs								
	$MAF_{1...20}$			Pairwise marker LD (D')	Marker-trait locus LD (Cramér's*V*)			
Flanking SNVs	0.1			0.0	0.0			
Map order	$H_{0,a}$: 10 flanking SNVs left – $\theta = 0.001$ – 3 test SNVs – $\theta = 0.001$ – 10 flanking SNVs right – $\theta = 0.5$ – trait locus							
	$H_{0,b}$, H_1: 10 flanking SNVs left – $\theta = 0.001$ – trait locus – $\theta = 0.0$ – 3 test SNVs – $\theta = 0.001$ – 10 flanking SNVs right							

Table 2. Population haplotype frequencies of the marker-trait locus haplotypes used for the simulations

TM1/2			
Population haplotype frequencies used for the simulations given as $SNV_1\|SNV_2\|SNV_3\|TL \in \{p_{h_0}, p_{h_1}, p_{h_2}, p_{h_3}, p_{h_4}, p_{h_5}, p_{h_6}, p_{h_7}, p_{h_8}, p_{h_9}, p_{h_{10}}, p_{h_{11}}, p_{h_{12}}, p_{h_{13}}, p_{h_{14}}, p_{h_{15}}\}$			
Frequencies	$H_{0,a}/H_{0,b}$	H_1	
		S1	S2	S3
$p_{h_0} = 0\|0\|0\|0$	0.010791	0.0101	0.0094	0.0059
$p_{h_1} = 0\|0\|0\|1$	0.000109	0.0008	0.0015	0.005
$p_{h_2} = 0\|0\|1\|0$	0.043659	0.04169	0.0411	0.044
$p_{h_3} = 0\|0\|1\|1$	0.000441	0.00241	0.003	0.0001
$p_{h_4} = 0\|1\|0\|0$	0.043659	0.043659	0.04409	0.044
$p_{h_5} = 0\|1\|0\|1$	0.000441	0.000441	0.00001	0.0001
$p_{h_6} = 0\|1\|1\|0$	0.000891	0.000891	0.00089	0.00089
$p_{h_7} = 0\|1\|1\|1$	0.000009	0.000009	0.00001	0.00001
$p_{h_8} = 1\|0\|0\|0$	0.043659	0.04169	0.04409	0.044
$p_{h_9} = 1\|0\|0\|1$	0.000441	0.00241	0.00001	0.0001
$p_{h_{10}} = 1\|0\|1\|0$	0.000891	0.00081	0.00089	0.00089
$p_{h_{11}} = 1\|0\|1\|1$	0.000009	0.00009	0.00001	0.00001
$p_{h_{12}} = 1\|1\|0\|0$	0.000891	0.000891	0.00089	0.00089
$p_{h_{13}} = 1\|1\|0\|1$	0.000009	0.000009	0.00001	0.00001
$p_{h_{14}} = 1\|1\|1\|0$	0.845559	0.850269	0.84865	0.84943
$p_{h_{15}} = 1\|1\|1\|1$	0.008541	0.003831	0.00545	0.00467

Table 3. Marginal haplotype frequencies of the marker-trait locus haplotypes for two SNVs (top) and a single (bottom) SNV and the trait locus, calculated from the haplotype frequencies for the marker-trait locus haplotypes for three SNVs and the trait locus used for the simulations (see Table 2)

TM1, TM2				
Marginal haplotype frequencies for the 2- and single-marker analyses (given as $SNV_1\|SNV_2\|TL$ and $SNV_2\|TL$, respectively). Values as derived from Table 2				
Frequencies	$H_{0,\,a}/H_{0,\,b}$	H_1		
		S1	S2	S3
2-marker				
$p_{h_0} = 0\|0\|0$	0.05445	0.05179	0.0505	0.0499
$p_{h_1} = 0\|0\|1$	0.00055	0.00321	0.0045	0.0051
$p_{h_2} = 0\|1\|0$	0.04455	0.04455	0.04498	0.04489
$p_{h_3} = 0\|1\|1$	0.00045	0.00045	0.00002	0.00011
$p_{h_4} = 1\|0\|0$	0.04455	0.0425	0.04498	0.04489
$p_{h_5} = 1\|0\|1$	0.00045	0.0025	0.00002	0.00011
$p_{h_6} = 1\|1\|0$	0.84645	0.85116	0.84954	0.85032
$p_{h_7} = 1\|1\|1$	0.00855	0.00384	0.00546	0.00468
Single-marker				
$p_{h_0} = 0\|0$	0.099	0.09429	0.09548	0.09479
$p_{h_1} = 0\|1$	0.001	0.00571	0.00452	0.00521
$p_{h_2} = 1\|0$	0.891	0.89571	0.89452	0.89521
$p_{h_3} = 1\|1$	0.009	0.00429	0.00548	0.00479

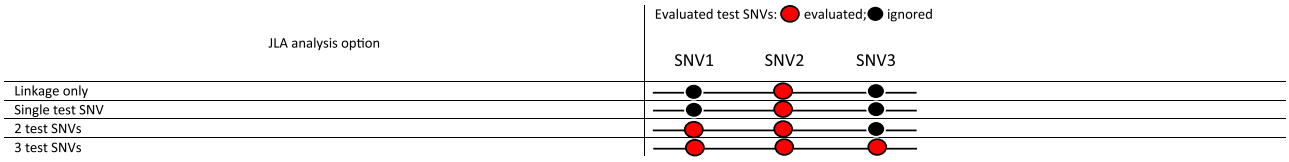
analysis, JLA analysis was performed using the left and the central test SNV (see also Table 4). The disease-allele frequency and penetrance restrictions were set to the default values ($p_m \leq 0.5; f_0 \leq f_{1,\text{pat}}, f_{1,\text{mat}} \leq f_2$). Imprinting analysis ($f_{1,\text{pat}} \neq f_{1,\text{mat}}$) was enabled for both trait models. In the case of GHM-MOD, the analysis was done using the following additional options: GHM option "maximization dense" for the optimization of the trait-model parameters using a dense grid of values, "calculate p value" to calculate *p* values (function "pmod") for the MOD score, "dimensions 5" to vary all five trait-model parameters simultaneously during the maximization. We compared type I error and power of the GHM-JLA tests with GHM-MOD and with the parsimonious JLA test implemented in the PSEUDOMARKER software [4, 63, 64] using the dominant and recessive PSEUDOMARKER models (PM-DOM, PM-REC) and with all other options set to their default values. PSEUDOMARKER-JLA tests were evaluated using the central test SNV, with *p* values reported as given by the program output. In addition, we compared linkage and association parameter estimates obtained from the JLA-MOD score with the values used for the simulations.

Analysis of FaPaCa Families

Pancreatic ductal adenocarcinoma (PDAC) is a challenging tumor entity with an increasing incidence and a dismal prognosis [98]. One of the greatest risk factors for developing PDAC is a positive family history [99]. When two or more first-degree relatives that do not fulfil the criteria for another inherited tumor syndrome have PDAC, this is called FaPaCa [99]. The German National Case Collection of FaPaCa, a tumor registry, was established as a screening program for an early detection of FPC and to further investigate its genetic and molecular basis [100, 101].

To demonstrate the applicability of the GHM-JLA analysis in practice, we analyzed pedigree data of the FaPaCa registry, consisting of genome-wide array-based genotypes that were obtained from peripheral blood samples for 193 individuals in 31 families. Family sizes ranged from triads to multigenerational complex pedigrees, with 409 individuals in total (overall genotyping rate: 47%). Patient records concerning pancreatic health status, which were gathered from family history or assessed during visits in the context of the FaPaCa screening program (see [101] and references therein for details), served as the basis for our phenotype definition. Affection status was set to "affected" if the individual had at least one of the following traits: pancreatic cancer (PC), pancreatic intraepithelial neoplasia-3 (PanIN-3), or intraductal papillary mucinous neoplasm with high-grade dysplasia. Screening of patients started 10 years before the youngest age of onset in the family or by the age of 40 (since 2016: 50) years, whichever occurred earlier. Over the years, several predisposing mutations have been identified mainly on the basis of co-occurring tumor types like breast cancer (BC) or colorectal carcinoma [101]. However, the genetic predisposition for many FPC families is still unknown [101]. Hence, in order to focus the gene discovery on those FPC families, for which the predisposing genetic background is unknown, we excluded families having at least one known predisposing genetic mutation in the gene set including *BRCA2*, *PALB2*, *CDNK2a*, *SUFU*, and *CHEK2* (see also [101, 102] for more details about the mutation screening panel). Individuals of an FPC family that solely had BC were marked as "unknown" because it has been shown that BC and PC have a common causal pathway, mediated, e.g., by *BRCA1/2* or *PALB2* mutations [103]. This procedure provides a compromise between setting these individuals to "unaffected," which is presumably wrong, or to "affected," which might have an unduly high impact on the analysis results. Individuals having patient records concerning pancreatic health status with no indication of PC, PanIN-3, intraductal papillary mucinous neoplasm with high-grade dysplasia, or BC, as assessed during the screening visits, were set to "unaffected." Despite differences in median ages, the age range of the first diagnosis of PC for affected in our final pedigree sample (37–86; median 65) was roughly comparable to the age range of the unaffected at their last screening visit (33–74; median 51). Because the definition of age-dependent thresholds and hence liability classes for developing PC in the familial context presents a complicated task and is beyond the scope of this paper, setting all individuals with a negative screening result to "unaffected," while setting unscreened individuals to "unknown," provides an acceptable working solution to map genes potentially involved in the complex FPC disease etiology. Genotyping was done using the Infinium Global Screening Array-24 v1.0 (GSAMD-24v1) from Illumina, which includes 700,078 variants. Genotype calling was performed using the Genome Studio 2.0 software (Illumina Inc. San Diego, California, USA). After calling with Genome Studio 2.0, a post-processing step of the data was done with zCall to refine the quality of rare variants [104]. The "Whole Genome Association Analysis Toolset" (PLINK 1.7 [105]) was used for the SNVs quality control. SNVs with a genotyping rate larger than 90% and not deviating from

Table 4. Overview of the test SNVs and JLA analysis options



Hardy-Weinberg equilibrium (significance threshold $p < 5{\cdot}10^{-6}$) were considered in the analysis. For the initial linkage scan using GHM, SNVs were chosen such that their MAF was larger than 25% and with pairwise LD between SNVs not exceeding 0.05 in terms of the squared correlation coefficient r^2 as calculated by PLINK. Errors in pedigree structure were identified using identical-by-descent analysis implemented in PLINK as well as the "scan pedigree" analysis option implemented in GHM. Relationships within and between pedigrees were investigated using the relationship estimation software packages KING [106] and TRUFFLE [107]. Genetic positions of the SNVs were obtained using the map file as provided by the manufacturer, which was based on the Genome Reference Consortium Human Build 37 (GRCh37).

The analysis procedure was as follows. First, we performed an initial standard linkage MOD score analysis using GHM with options "modcalc global," "imprinting on," "allfreq restriction on," "penetrance restriction on," "max bits 20," "maximization dense," "dimensions 5," and "increment step 2." Then, chromosomes with a MOD score larger than 3.0 were chosen for JLA analysis. To this end, the SNV lying next to the maximum linkage signal was used as the central test SNV in JLA analysis. Additional SNVs on either side of the central test SNV were added to the dataset, such that JLA analysis could be performed with a single, two, and three test marker(s) forming the marker-trait locus haplotype. The additionally added SNVs also had to pass the abovementioned quality control; however, the MAF had to be at least 5% and the pairwise LD in terms of r^2 between each test SNV and the two flanking linkage markers was not allowed to exceed 0.1, which should still eliminate the risk of inflated multipoint linkage scores when parental genotypes are not available [45, 91]. Because most of the parental genotypes of the FaPaCa families were not available, pedigrees were pruned for JLA analysis to keep the computations still feasible. Specifically, pedigrees were pruned such that no pedigree had more than two untyped founders, except for half-sibs, which were allowed to have three untyped founders. As it was for the initial linkage scan, the disease-allele frequency and penetrance restrictions were set to the default values ($p_m \leq 0.5$; $f_0 \leq f_{1,\text{pat}}, f_{1,\text{ mat}} \leq f_2$), and imprinting analysis ($f_{1,\text{ pat}} \neq f_{1,\text{ mat}}$) was enabled. Empiric p values were obtained using 999 simulated replicates. Due to the exploratory nature of the analysis, p values ≤0.05 were considered statistically significant.

Results

The results section is structured as follows. In the first part, we present the results of the simulated scenarios with a focus on type I error rate and power of the GHM-JLA analyses as well as the empiric distribution of the JLA-MOD score. We also demonstrate the validity of the new GHM-JLA simulation procedure to obtain an empiric p value for the JLA test. Furthermore, we briefly discuss the accuracy of the estimated trait-model parameters as well as the estimated haplotype frequencies obtained from the GHM-JLA analyses. In the second part, we compare the results obtained from our GHM-JLA method with those obtained from the PSEUDOMARKER-JLA analyses with respect to type I error and power. In the final part, we present the results of the real data application, i.e., the GHM-JLA analysis of the FaPaCa families.

Type I Error, Power, and Parameter Estimation

Simulation Scenario $H_{0,\,a}$: No Linkage, No Association

The results for the GHM-MOD and GHM-JLA analyses for the datasets simulated under the null hypothesis of no linkage and no association can be found in Tables 5 and 6 as well as in online supplementary Table 1 (upper part) (for all online suppl. material, see https://doi.org/10.1159/000535840). As can be deduced from Table 5, the type I error rates of the linkage as well as all JLA tests corresponded well to their nominal significance level of 5%. With regard to the results in Table 6, p values for the linkage test were comparable, irrespective of the method to generate replicates to obtain empiric p values, i.e., either using the GHM function "pmod" or the GHM-JLA replicates. This can be interpreted as a confirmation of the validity of our new JLA simulation procedure to generate replicates under the null hypothesis of no linkage and no association. In the same line, the obviously low trait-model parameter estimation performance of the JLA tests did not differ between the original datasets and the JLA replicates (online suppl. Table 1).

The results regarding the haplotype frequencies for the single-, 2-, and 3-SNV haplotypes estimated using the EM algorithm can be found in online Supplementary Figure 1 (left column). As can be deduced from online supplementary Figure 1, the estimated haplotype frequencies were in good accordance with the simulated values across

Table 5. Overview of type I error rate and power of the GHM-linkage and GHM-JLA tests for the simulated scenarios

GHM analysis option	Simulation scenario								
	$H_{0,\,a}$	$H_{0,\,b}$: TM1	$H_{0,\,b}$: TM2	H_1: TM1, S1	H_1: TM1, S2	H_1: TM1, S3	H_1: TM2, S1	H_1: TM2, S2	H_1: TM2, S3
Linkage only*	0.054	0.487	0.687	0.480	0.451	0.495	0.667	0.683	0.686
1-SNV test marker	0.049	0.365	0.584	0.898	0.751	0.854	0.972	0.933	0.957
2-SNV test markers	0.055	0.291	0.478	0.842	0.820	0.886	0.952	0.940	0.959
3-SNV test markers	0.053	0.276	0.452	0.772	0.766	0.976	0.912	0.920	0.983

*Values averaged based on the three corresponding results in column "PMOD" in Table 6.

all JLA test marker scenarios. With respect to the haplotype frequencies of the test SNV alleles and the alleles at the disease locus (online suppl. Fig. 1, right column), the frequencies deviated from the simulated values due to the overestimation of the disease-allele frequency, given no linkage and hence no power for the JLA tests (see also online suppl. Table 1, top).

Simulation Scenario $H_{0,\,b}$: Linkage, No Association

The results for the GHM-JLA analyses for the datasets simulated under the hypothesis of linkage and no association can be found in Tables 5 and 6 as well as in online supplementary Table 1 (middle and lower part). As to the trait model TM1, the linkage test showed higher power (0.487) than the JLA tests (0.365, 0.291, and 0.276 for the analyses using one, two, or three test SNVs, respectively). This is due to an increased effective number of *df* for the JLA tests as compared to the linkage test. In the same line, the power of the JLA tests decreased with an increasing number of test SNVs and hence parameters for the MOD score. The same held true for the trait model TM2, albeit the power was generally higher for all tests as compared to TM1. This is because the linkage and all JLA tests allowed for imprinting models, which lead to an increased power if imprinting is really present, as it is for TM2.

With regard to Table 6, *p* values for the linkage test were comparable, irrespective of the method to generate replicates to obtain empiric *p* values. This was in line with the results obtained under $H_{0,\,a}$ (see above).

The estimation accuracy of individual trait-model parameters was generally low for both trait models (see online suppl. Table 1), which means that estimates and standard deviations did not differ much from those obtained from the corresponding $H_{0,\,a}$ replicates. This is mainly due to the fact that the power of the JLA tests was rather low (0.276–0.365 for TM1 and 0.452–0.584 for TM2, see Table 5). Yet, the LD parameter *V*, the penocopy rate f_0, and the heterozygote penetrance of the imprinted sex together with the imprinting index *I* were estimated with increased accuracy as compared to the null hypothesis replicates.

The results for the EM-estimated haplotype frequencies of all JLA test marker sets can be found in online supplementary Figure 2 (left column) for TM1 and in online supplementary Figure 3 (left column) for TM2, which were in good accordance with the simulated values for both trait models. The corresponding haplotype frequencies of the test SNV alleles and the alleles at the disease locus showed an improved accuracy compared to those obtained under $H_{0,\,a}$ due to an improved estimation accuracy of the disease-allele frequency. This was especially true for TM2 due to an increased power for the JLA tests compared to TM1 (see also Table 5; online suppl. Table 1, middle and bottom).

Simulation Scenario H_1: Linkage, Association

TM1. The results for the GHM-JLA analyses for the datasets simulated under the hypothesis of linkage and association and using trait model TM1 can be found in Tables 5 and 6 as well as in online supplementary Table 2. As can be seen from Table 5, the power of the linkage test did not substantially change compared to the $H_{0,\,b}$ scenarios, irrespective of the extent of LD (S1, S2, or S3). With respect to scenario S1, the power of the JLA tests was higher than the power of the linkage test (0.48) and decreased with an increasing number of test SNVs (0.898, 0.842, and 0.772 for the JLA test using one, two, or three test SNVs, respectively). As to scenario S2, the JLA test with two test SNVs showed higher power than the linkage test and the tests with one or three test SNVs (0.82 vs. 0.451, 0.751, and 0.766, respectively). With regard to scenario S3, the JLA test with three test SNVs showed the highest power of all tests (0.976 vs. 0.495, 0.854, and 0.886 for the linkage test and the JLA tests using one or two test

SNVs, respectively). With regard to Table 6, p values for the linkage test were comparable, irrespective of the method to generate replicates to obtain empiric p values. This was in line with the results obtained under $H_{0,\,a}$ and $H_{0,\,b}$ (see above).

As can be deduced from online supplementary Table 2, the parameter estimation accuracy generally improved due to the increased power of the JLA tests under H_1 as compared to $H_{0,\,b}$. Specifically, estimates for the disease-allele frequency p_m, the phenocopy rate f_0, the imprinting index I, and the LD parameter V showed improved accuracy as compared to the $H_{0,\,b}$ scenario. Interestingly, parameter estimation performance did not substantially differ between the three JLA tests.

The results for the EM-estimated haplotype frequencies of all JLA test marker sets for the LD scenarios S1, S2, and S3 can be found in online supplementary Figures 4–6 (left columns), respectively. In contrast to the results under $H_{0,\,a}$ and $H_{0,\,b}$, the corresponding haplotype frequencies slightly deviated from the simulated values, which is likely due to marker-dependent ascertainment/sampling of pedigrees under H_1. This way, the haplotype frequency distribution in the ascertained pedigree sample does no longer correspond to the population haplotype frequency distribution, although the difference can be mitigated by including more healthy controls [108]. The results of the corresponding haplotype frequencies of the test SNV alleles and the alleles at the disease locus showed an improved accuracy compared to those obtained under $H_{0,\,a}$ and $H_{0,\,b}$ due to the higher power of the JLA tests under H_1 (online suppl. Fig. 4–6, right columns).

TM2. The results for the GHM-JLA analyses for the datasets simulated under the hypothesis of linkage and association and using trait model TM2 can be found in Tables 5 and 6 as well as in online supplementary Table 3. As can be seen from Table 5, the power of the linkage test did not substantially change compared to the corresponding $H_{0,\,b}$ scenarios, irrespective of the extent of LD (S1, S2, or S3). With respect to scenario S1, the power of all JLA tests was higher than the power of the linkage test (0.667) and decreased with an increasing number of test SNVs (0.972, 0.952, and 0.912 for the analyses using one, two, or three test SNVs, respectively). As to scenario S2, the JLA analysis with two test SNVs showed higher power than the linkage test and the tests with one or three test SNVs (0.94 vs. 0.683, 0.933, and 0.92, respectively). With regard to scenario S3, the JLA test with three test SNVs showed the highest power of all tests (0.983 vs. 0.686, 0.957, and 0.959 for the linkage test and the JLA tests using one or two test SNVs, respectively). With regard to Table 6, p values for the linkage test were comparable,

Table 6. Comparison of type I error rate and power for the GHM-linkage test using either the GHM analysis option "pmod" (PMOD) or the JLA replicates (JLA) to calculate empiric p values

GHM-linkage test	Simulation scenario																	
JLA replicates generated using JLA option	$H_{0,\,a}$		$H_{0,\,b}$: TM1		$H_{0,\,b}$: TM2		H_1: TM1, S1		H_1: TM1, S2		H_1: TM1, S3		H_1: TM2, S1		H_1: TM2, S2		H_1: TM2, S3	
	PMOD	JLA	PMOD	JLA	PMOD	JLA	PMOD	JLA	PMOD	JLA	PMOD	JLA	PMOD	JLA	PMOD	JLA	PMOD	JLA
1-SNV test marker	0.057	0.057	0.488	0.484	0.686	0.686	0.479	0.482	0.449	0.445	0.494	0.491	0.669	0.669	0.682	0.688	0.683	0.685
2-SNV test markers	0.052	0.052	0.485	0.480	0.687	0.686	0.477	0.477	0.450	0.449	0.496	0.496	0.665	0.669	0.687	0.680	0.689	0.677
3-SNV test markers	0.054	0.052	0.487	0.484	0.687	0.686	0.483	0.477	0.453	0.452	0.495	0.488	0.666	0.663	0.681	0.683	0.686	0.677

irrespective of the method to generate replicates to obtain empiric *p* values. This was in line with the results obtained under $H_{0,\ a}$, $H_{0,\ b}$, and H_1 with TM1 (see above).

With regard to online supplementary Table 3, the parameter estimation accuracy generally improved due to the increased power of the JLA tests under H_1 as compared to $H_{0,\ b}$. Specifically, estimates for the disease-allele frequency, the phenocopy rate, the imprinting index, and the LD parameter showed improved accuracy as compared to the $H_{0,\ b}$ scenario. In line with the results for TM1, parameter estimation performance did not substantially differ between the three JLA tests. The difference in power between the three JLA tests was smaller across all LD scenarios as compared to the results obtained for TM1. The generally higher power for the TM2 analyses compared to the TM1 analyses is due to the fact that for TM1 imprinting is absent, but accounted for in the analyses, while imprinting is in fact present for TM2.

The results for the EM-estimated haplotype frequencies of all JLA test marker sets for the LD scenarios S1, S2, and S3 can be found in online supplementary Figures 7–9 (left columns), respectively. As it was for TM1, the corresponding haplotype frequencies slightly deviated from the simulated values compared to the results under $H_{0,\ a}$ and $H_{0,\ b}$, which is likely due to marker-dependent ascertainment/sampling of pedigrees under H_1 (see explanation above). The results of the corresponding haplotype frequencies of the test SNV alleles and the alleles at the disease locus showed an improved accuracy compared to those obtained under $H_{0,\ a}$, $H_{0,\ b}$, and H_1 with TM1 due to the higher power of the JLA tests under H_1 with TM2 (online suppl. Fig. 7–9, right columns).

JLA-MOD Score Distribution

The empiric distributions of the JLA-MOD score based on one, two, and three test SNVs and for all investigated simulation scenarios can be found in Figures 1–3, showing the results for $H_{0,\ a}$ and $H_{0,\ b}$, for H_1 and TM1, and for H_1 and TM2, respectively. As to $H_{0,\ a}$ and $H_{0,\ b}$ (Fig. 1), the empiric distribution of the JLA-MOD score was shifted toward larger values with an increasing number of test SNVs. This is because of the increasing number of effective *df* with an increasing number of test SNVs in the JLA test. The corresponding histograms indicated that the COBYLA optimization algorithm used in GHM 4 worked properly, meaning that artificial patterns in the empiric distributions like, e.g., excess point masses around 0.0 could not be observed. In accordance with the power values in Table 5, the empiric distributions for the JLA-MOD scores of the original SLINK datasets simulated under H_1 (Fig. 2; 3) were all shifted toward higher values as compared to the distributions obtained under $H_{0,\ a}$ and $H_{0,\ b}$ (Fig. 1), with even higher values for TM2 as compared to TM1. Despite a few larger outlying values, the empiric JLA-MOD score distributions all showed an approximately continuous, unimodal curve with no obvious aberrant pattern, which would otherwise indicate problems during the optimization process of the JLA-MOD score calculation.

Comparison with PSEUDOMARKER

The results of the PSEUDOMARKER analyses are summarized in Table 7. With respect to $H_{0,\ a}$, the quality of the asymptotic distributions for both PSEUDOMARKER models PM-DOM and PM-REC was moderate (true type I errors 0.0715 and 0.0744 for PM-DOM and PM-REC, respectively, assuming a nominal type I error rate of 0.05). Under $H_{0,\ b}$, the power did not exceed 0.18 for both PM-DOM and PM-REC as well as for both trait models TM1 and TM2 (Table 7), whereas the power ranged from 0.276 to 0.584 using the GHM-JLA tests (Table 5). Under H_1 and for TM1, the power ranged from 0.643 to 0.822 for PM-DOM and from 0.528 to 0.721 for PM-REC (Table 7). The highest power was detected for the S1 LD scenario, followed by S3. The power was consistently higher for PM-DOM as compared to PM-REC. The corresponding power values for the GHM-JLA tests were consistently higher for the S2 and S3 scenarios. In the case of the S1 scenario, PM-DOM showed higher power than the GHM-JLA test using 3 SNVs, which showed the lowest power among the GHM-JLA tests for this scenario (0.822 vs. 0.772, respectively, see Tables 5, 7). Under H_1 and for TM2, the power ranged from 0.68 to 0.789 for PM-DOM and from 0.621 to 0.782 for PM-REC (Table 7). Again, the highest power was detected for the S1 LD scenario, followed by S3. The power was again consistently higher for PM-DOM as compared to PM-REC, and it was mostly higher as compared to the corresponding results for TM1. The corresponding power values for the GHM-JLA tests were consistently higher for all LD scenarios. With regard to the S2 scenario, the GHM linkage-only test even outperformed the PSEUDOMARKER-JLA test (GHM linkage-only: 0.683 vs. PM-DOM: 0.680 and PM-REC: 0.621). A graphical overview of all the type I error and power values for both the PSEUDOMARKER and GHM-JLA analyses is given in Figure 4.

Analysis of FaPaCa Families

Identical-by-descent analyses of the FaPaCa families led to the exclusion of a duplicated individual. The relationship estimation algorithms did not find any significant deviation from the relationships given in the

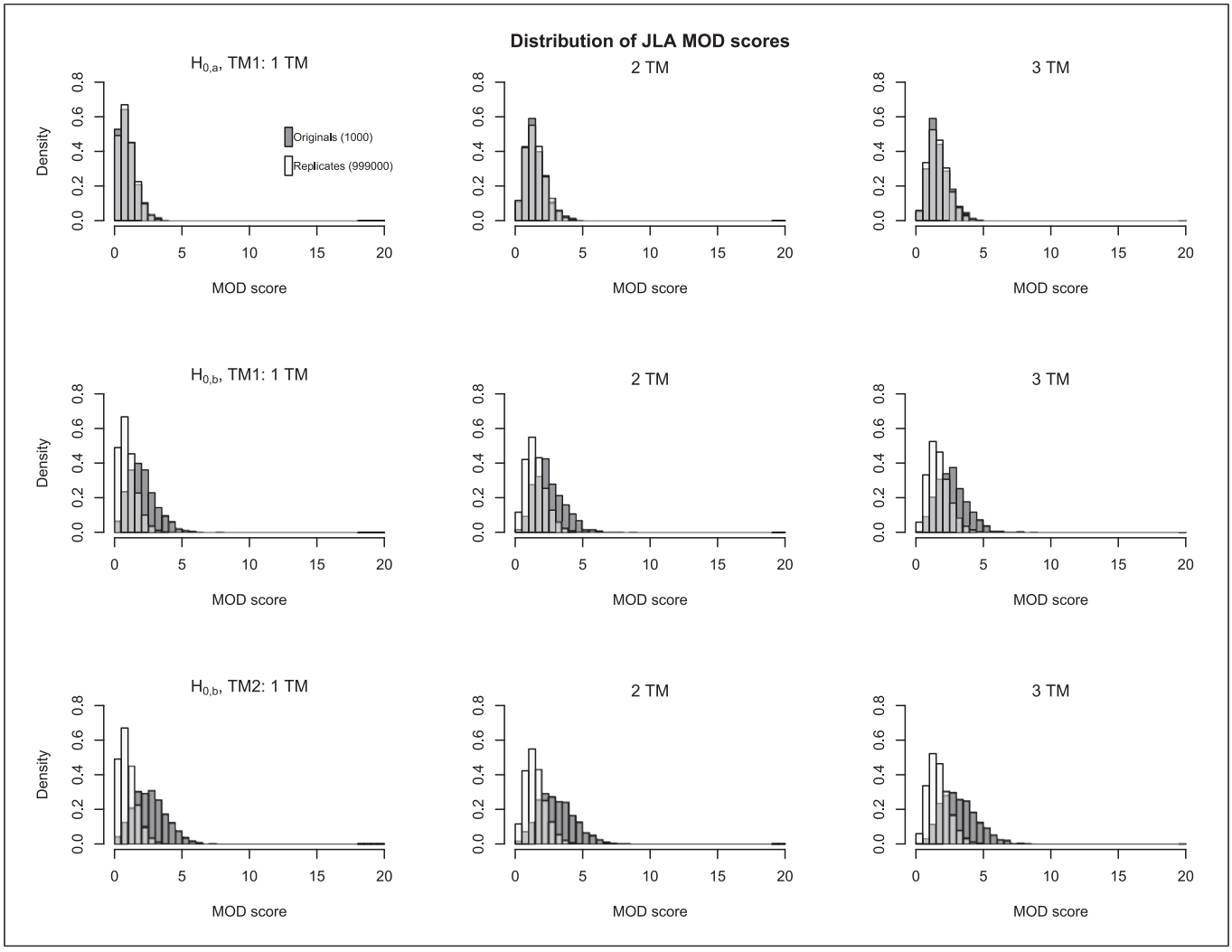


Fig. 1. Depiction of the empiric distributions of JLA-MOD scores for data simulated under the hypothesis of no linkage, no association (row 1, depiction only for trait model TM1) and linkage, no association (row 2 for TM1, row 3 for TM2). The bars of the JLA-MOD scores of the "original" simulated SLINK datasets are colored in dark-gray; the bars of the simulated GHM replicates are colored in white, overlapping areas are colored in light-gray. For more information about the simulation scenarios, see Table 1.

pedigree tree and those estimated using the genetic data. Further, no interrelatedness between pedigrees could be observed. In total, the final sample consisted of 262 individuals in 22 pedigrees, with 78 affected, 47 unaffected, and 137 unknowns. After the initial standard linkage MOD score analysis on all autosomes, chromosome 22 (MOD score: 3.09 near marker rs5771131 within the *TTLL8* gene on 22q13.33) was further investigated using JLA analysis. To refine the candidate region for JLA analysis, we repeated the GHM-linkage scan for chromosome 22, but now with the option "modcalc single" to obtain best-fitting trait models for every investigated genetic position, which allows a better evaluation of the width of the linkage signal than the "modcalc global" option (see online suppl. Fig. 10). Because the candidate region showed distinctive sex-specific recombination fractions, we repeated the linkage scan using the sex-specific genetic distances as given in the Rutgers map v.3 [109] and assuming the Haldane map function, which did not significantly change the results. We then chose four additional SNVs in the vicinity of rs5771131 and encompassing the two nearby candidate genes *IL17REL* and *PIM3*, according to our criteria given above in the Methods section. The results of the ensuing JLA analysis

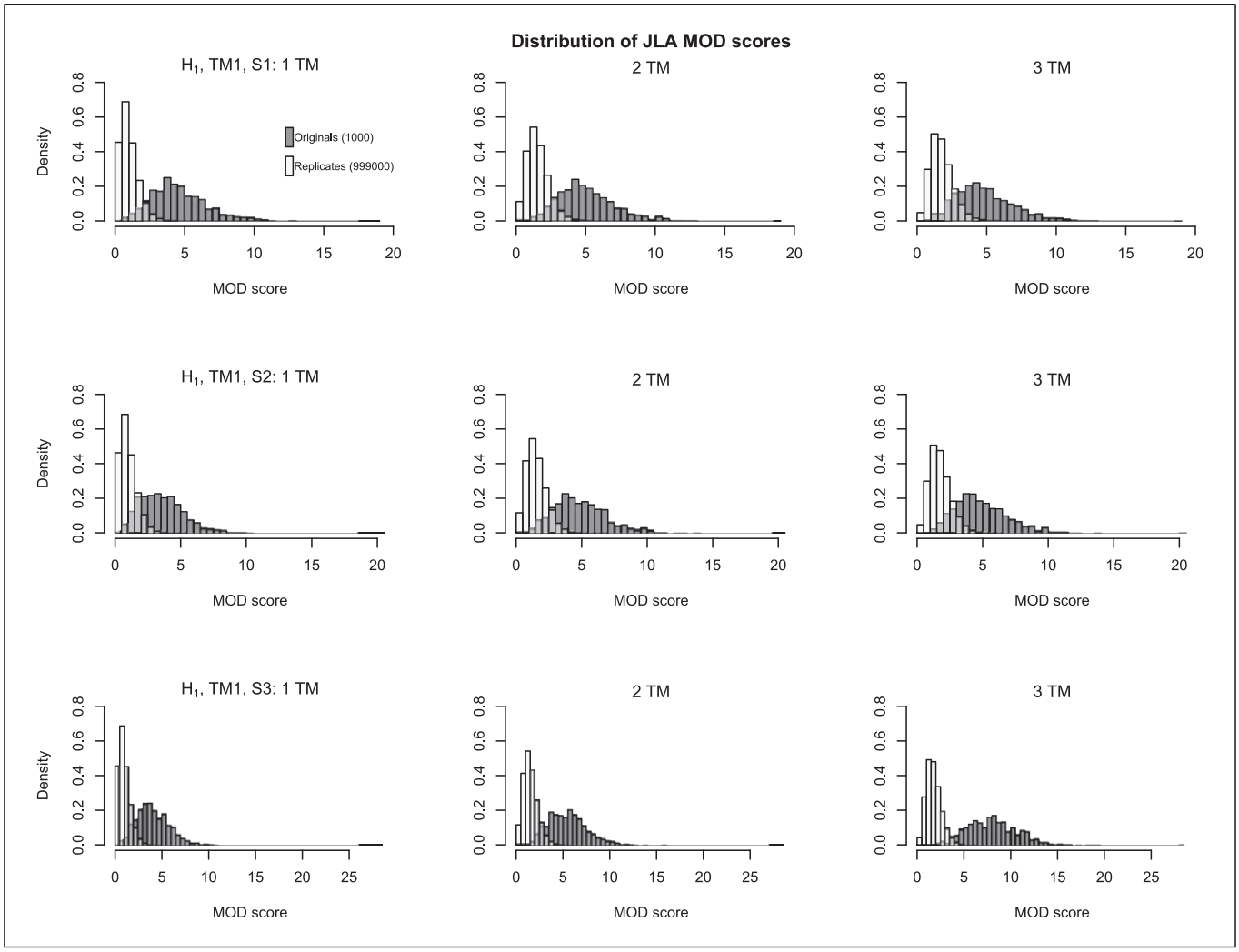


Fig. 2. Depiction of the empiric distributions of JLA-MOD scores for data simulated under the hypothesis of linkage and association for trait model TM1 and various LD patterns (row 1: S1; row 2: S2; row 3: S3). For more details, see Figure 1.

can be found in Table 8. In summary, significant results were obtained for one single test SNV, two sets of two test SNVs, and four sets of three test SNVs, all with an imprinting index pointing toward maternal imprinting (Table 8). Remarkably, at least one of the neighboring markers rs5771069 and rs137878 was present in every significant test set.

Discussion

In this work, we present an extension to the GENEHUNTER-MODSCORE software package [16–19] that allows a JLA analysis using pedigrees, triads, and unrelated individuals. The implementation to perform a JLA analysis using MOD scores has been missing so far. Our new GHM version 4 now closes this gap. In GHM 4, association is modeled using haplotype frequencies for up to three diallelic test markers and a diallelic trait locus. In addition, we also provide an integrated simulation routine to calculate empiric p values for the JLA test.

We demonstrated by simulations that a JLA analysis based on MOD scores can substantially increase power as compared to the corresponding linkage-only test (Table 5). This observation was in accordance with the statement mentioned earlier, saying that a JLA analysis can substantially increase mapping accuracy and power because it makes use of both family and population

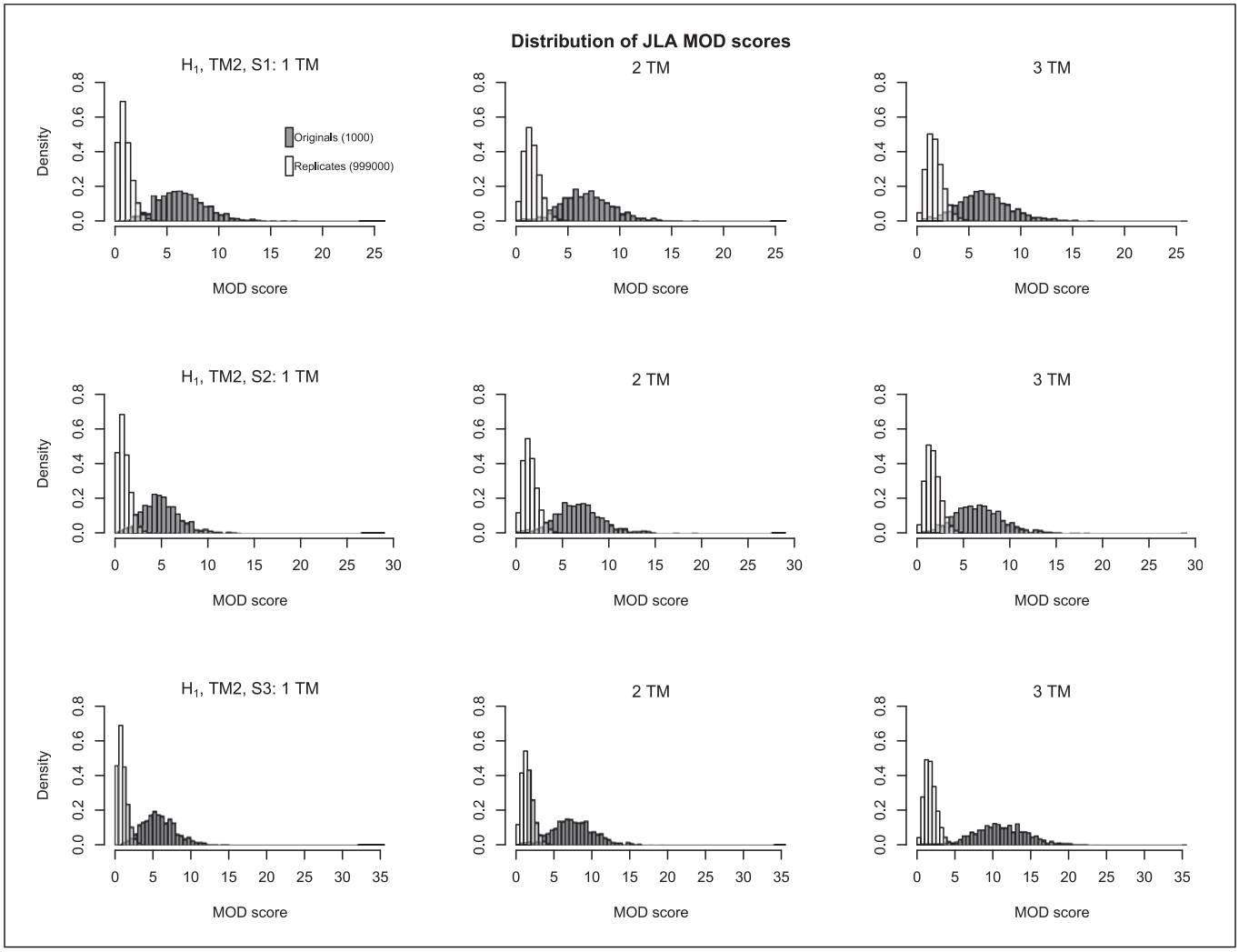


Fig. 3. Depiction of the empiric distributions of JLA-MOD scores for data simulated under the hypothesis of linkage and association for trait model TM2 and various LD patterns (row 1: S1; row 2: S2; row 3: S3). For more details, see Figure 1.

Table 7. Overview of type I error rate and power of the PSEUDOMARKER-JLA tests for the simulated scenarios as reported by the PSEUDOMARKER software

PSEUDOMARKER analysis option	Simulation scenario								
	$H_{0,\ a}$*	$H_{0,\ b}$: TM1	$H_{0,\ b}$: TM2	H_1: TM1, S1	H_1: TM1, S2	H_1: TM1, S3	H_1: TM2, S1	H_1: TM2, S2	H_1: TM2, S3
PM-DOM	0.0715	0.160	0.178	0.822	0.643	0.766	0.789	0.680	0.754
PM-REC	0.0744	0.136	0.157	0.721	0.528	0.675	0.782	0.621	0.733

PM-DOM and PM-REC correspond to using the dominant and recessive pseudomarker model in the JLA analysis, respectively. The PSEUDOMARKER-JLA tests are supposed to asymptotically follow a 50-50 mixture of χ^2_1 and χ^2_2 distributions in the case of a diallelic test marker. *Based on 10,000 SLINK replicates.

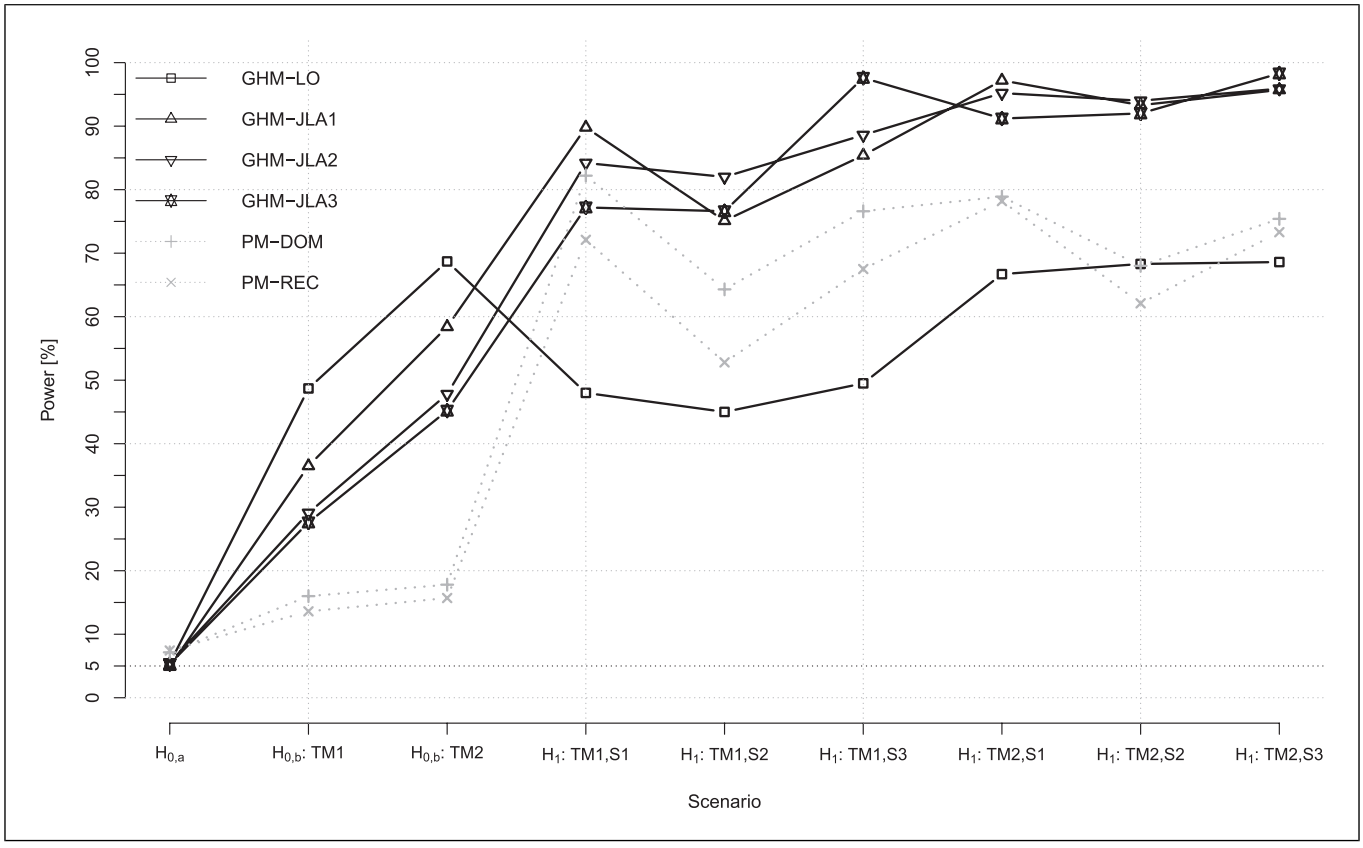


Fig. 4. Depiction of type I error and power values for the six investigated test statistics and analysis options: GHM-LO: GHM linkage-only MOD score; GHM-JLA1: GHM-JLA-MOD score using one test SNV for the analysis; GHM-JLA2: GHM-JLA-MOD score using two test SNVs for the analysis; GHM-JLA3: GHM-JLA-MOD score using three test SNVs for the analysis; PM-DOM: PSEUDOMARKER analysis assuming a dominant model for the analysis; PM-REC: PSEUDOMARKER analysis assuming a recessive model for the analysis. For more information about the simulation scenarios, see Table 1.

information [4, 5]. Moreover, we showed that there are LD scenarios, for which either the 2- or 3-marker JLA tests can be more powerful than the corresponding single-marker test, which confirms another statement mentioned earlier, saying that haplotype-based association methods can outperform single-marker analyses [71], especially when the LD between the investigated test markers and the trait locus is rather weak [73].

The problem as to whether either single-marker or haplotype-based JLA tests are generally more powerful is hard to tackle. Of course, an already high degree of LD between alleles at a single marker and the alleles at the trait locus renders the extra LD information gathered from additional markers less important. However, apart from LD information, additional test markers can contribute valuable linkage information for the JLA test when there is reduced linkage information at a single test marker locus. Furthermore, it is conceivable that LD can likely be modeled more efficiently using haplotype-based approaches when there are several independent disease-associated SNVs in the same LD region [71]. Generally, whether single-marker or multi-marker haplotypes are more suitable in a JLA analysis depends on the disease etiology as a function of the mode of inheritance (number of disease loci, disease-allele frequencies, penetrances) and the population history defining the LD block.

The ability to estimate trait-model parameters using MOD score analysis has been thoroughly discussed in the literature [12–15, 70]. In the case of a JLA analysis, trait-model parameter estimates obtained from a MOD score analysis are argued to be trivially biased [14, 70]. In this work, however, we did not quantify this bias in detail because the JLA extension of the MOD score with several additional LD parameters makes the corresponding parameter estimation less efficient, and the quantification of the bias becomes unfeasible. Nevertheless, the parameter

Table 8. Results of the JLA analyses of the FaPaCa pedigrees using GHM. Chromosome 22 showing a MOD score for the GHM-linkage test larger than 3.0 was selected for JLA analysis

Chromosome 22: Nearest protein-coding genes	SNV1	SNV2	SNV3	LD	Imprinting index	JLA-MOD score	*p* value*
TTLL8 IL17REL PIM-3	rs28634968			0.013	1.0	1.72	0.178
	rs5771069			0.311	0.91	2.62	**0.039**
	rs137878			0.155	0.0	1.01	0.507
	rs5771131			0.008	1.0	2.08	0.100
	rs7290681			0.033	1.0	1.50	0.243
	rs28634968	rs5771069		0.329	1.0	2.88	0.078
	rs28634968	rs137878		0.231	1.0	2.03	0.241
	rs28634968	rs5771131		0.329	0.40	1.71	0.399
	rs28634968	rs7290681		0.368	0.0	1.50	0.431
	rs5771069	rs137878		0.521	0.97	3.65	**0.025**
	rs5771069	rs5771131		0.314	0.92	2.96	0.099
	rs5771069	rs7290681		0.474	1.0	3.13	0.053
	rs137878	rs5771131		0.427	0.70	3.70	**0.027**
	rs137878	rs7290681		0.704	−0.35	1.66	0.428
	rs5771131	rs7290681		0.17	1.0	2.27	0.219
	rs28634968	rs5771069	rs137878	0.573	1.0	4.48	**0.023**
	rs28634968	rs5771069	rs5771131	0.298	0.93	3.34	0.227
	rs28634968	rs5771069	rs7290681	0.514	0.0	3.33	0.170
	rs28634968	rs137878	rs5771131	0.408	0.60	4.38	**0.040**
	rs28634968	rs137878	rs7290681	0.377	0.0	2.51	0.393
	rs28634968	rs5771131	rs7290681	0.268	1.0	2.44	0.460
	rs5771069	rs137878	rs5771131	0.537	0.78	4.50	**0.031**
	rs5771069	rs137878	rs7290681	0.585	0.91	4.12	0.062
	rs5771069	rs5771131	rs7290681	0.411	0.88	3.79	0.176
	rs137878	rs5771131	rs7290681	0.463	0.57	4.91	**0.029**

LD is given in terms of Cramér's *V*. *Based on 999 GHM replicates. Bold values are statistically significant, $p < 0.05$.

estimates obtained from the JLA-MOD score analyses in our simulation study under the alternative hypothesis of linkage and association often contained at least some degree of information as opposed to those obtained for the replicates under the null hypothesis of no linkage and no association (online suppl. Tables 1–3). Furthermore, the estimates for the imprinting index were in good accordance with the simulated values, which means that a JLA-MOD score analysis can also be used to quantify the imprinting effect as it is possible with the linkage-only MOD score [69].

We compared our MOD score JLA test to another commonly used parsimonious JLA test as implemented in the PSEUDOMARKER software package [4, 63, 64]. For the two scenarios under linkage but no LD as well as for five out of six scenarios with linkage and LD, the MOD score JLA tests showed consistently higher power than the PSEUDOMARKER tests. In the LD scenario S1, in which the single-marker MOD score JLA test outperformed the 2- and 3-marker MOD score JLA tests and which was simulated under no imprinting (TM1), the PSEUDOMARKER test assuming a dominant model showed higher power than the three-marker MOD score JLA test (Fig. 4).

Although limited to moderately sized pedigrees, GHM can efficiently calculate MOD scores by the use of many markers in a multipoint setting. The multipoint calculation enables the MOD score JLA test to incorporate flanking marker information, which can substantially increase power as compared to a twopoint approach as we have shown in this work. This is because, in the twopoint setting, all linkage and LD information is gathered only from the single test marker. Admittedly, the twopoint PSEUDOMARKER tests are capable of analyzing markers with more than two alleles, which can entail higher information content at the test marker locus; however, the availability of highly polymorphic markers is often limited in current research projects. Notwithstanding, the successful applicability of PSEUDOMARKER-JLA tests to mixed pedigree data including larger multigenerational pedigrees is undoubted (see, e.g., [110]).

The analysis of the FaPaCa data led to the identification of a novel candidate region for mutation analysis in FPC families on chromosome 22q13.33. The long arm of chromosome 22 has long been suspected to harbor genetic loci involved in the etiology of PDAC [111] and endocrine pancreatic tumors [112] using loss of heterozygosity mapping; however, the precise genetic loci involved in the etiology of PC on 22q are still unknown. Our newly discovered region encompasses the locus of the proto-oncogene PIM3, a serine/threonine-protein kinase showing enhanced expression in human PC cells [113], and the cytokine receptor IL17REL, which was found to be associated with inflammatory bowel disease [114] being a potential risk factor for PDAC [115]. Interestingly, the candidate region showed a considerable paternal expression pattern, corresponding to maternal imprinting. Data on imprinted genes in the context of PDAC are rare [116], but in light of the longer male genetic map in this region, the observed maternal imprinting – at least to some degree – might stem from a true signal rather than from confounding [117].

With GHM 4, it is now possible to jointly analyze mixtures of pedigrees and unrelated individuals in a joint test for linkage and association using up to three diallelic test markers. The computational burden involved in MOD score JLA analysis is substantial; however, calculations are still feasible on most present-day computing clusters. To save elapsed real time for the computations, GHM 4 offers an option to compute empiric p values in parallel. Moreover, GHM 4 offers the possibility to estimate haplotype frequencies by the use of the EM algorithm. We have demonstrated by simulations that the MOD score JLA test has good power under various linkage and LD scenarios and has the potential to characterize the disease gene to some extent, especially when imprinting is present. The MOD score JLA tests all keep the specified type I error level using a verified integrated simulation procedure, which can automatically be run in parallel. GHM 4 thus provides a valuable and powerful genetic analysis toolbox, unifying MOD score linkage with haplotype-based association analysis.

Acknowledgments

Parts of this research were conducted using the supercomputer Mogon 2 and advisory services offered by Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP (Alliance for High Performance Computing in Rhineland Palatinate, www.ahrp.info) and the Gauss Alliance e.V. The authors gratefully acknowledge the computing time granted on the supercomputer Mogon 2 at Johannes Gutenberg University Mainz (hpc.uni-mainz.de).

We thank Clemens Baumbach for his advice concerning programming details and for constantly fruitful discussions. Finally, we would like to thank the reviewers for their thoughtful comments that helped us improve the manuscript.

Statement of Ethics

The FaPaCa registry, including the genetic analyses and the screening program, was approved by the Ethics Committee of the Philipps-University of Marburg (36/1997, last amendment 9/2010). All participants provided written informed consent.

Conflict of Interest Statement

The authors declare that they have no competing interests.

Funding Sources

This work was supported by grant Str643/6-1 of the Deutsche Forschungsgemeinschaft (German Research Foundation). This work was also supported by a grant from the Wilhelm Sander-Stiftung (No. 2018.022.1) and a generous donation from the GAUFF-Foundation. Further, this research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

Author Contributions

Markus Brugger developed and implemented the new version of the GENEHUNTER-MODSCORE software, designed and performed the simulation study as well as the analysis of the FaPaCa data, and drafted the initial manuscript. Manuel Lutz and Martina Müller-Nurasyid were responsible for curating the FaPaCa data, involving quality control analyses, data preparation, and documentation. Peter Lichtner was responsible for genotyping the FaPaCa samples. Elvira Matthäi, Emily P. Slater, and Detlef K. Bartsch have been responsible for the long-term FaPaCa study management and data collection; they gave significant advice with regard to study design, phenotype definition, and suitable inclusion criteria for the FaPaCa data analysis. Konstantin Strauch planned and designed the new version of the GENEHUNTER-MODSCORE software, contributed substantially to the simulation and data analysis designs, and initiated and coordinated the project. All authors contributed to the article and approved the submitted version.

Data Availability Statement

The new version GENEHUNTER-MODSCORE 4 can be freely downloaded from our website: https://www.unimedizin-mainz.de/imbei/biometriegenomische-statistik-und-bioinformatik/software.html. Files and scripts used to generate the datasets for the simulation

study can readily be obtained upon request from the corresponding author.

The individual-level data of the FaPaCa study are not publicly available because the data contain sensitive patient data, which underlie data protection rules. This is in accordance with the local ethic vote and the regulations of the FaPaCa registry. Patients' characteristics are available upon request from the FaPaCa study registry (contact information: fapaca@med.uni-marburg.de).

References

1 Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33(Suppl l):228–37.
2 Terwilliger JD. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet. 1995;56(3):777–87.
3 Graham J, Thompson EA. Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am J Hum Genet. 1998;63(5):1517–30.
4 Göring HH, Terwilliger JD. 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. 2000;66(4):1310–27.
5 Lou XY, Ma JZ, Yang MC, Zhu J, Liu PY, Deng HW, et al. Improvement of mapping accuracy by unifying linkage and association analysis. Genetics. 2006;172(1):647–61.
6 Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet. 2015;16(5):275–84.
7 Knapp M, Seuchter SA, Baur MP. Linkage analysis in nuclear families. 2: Relationship between affected sib-pair tests and lod score analysis. Hum Hered. 1994;44(1):44–51.
8 Strauch K. MOD-score analysis with simple pedigrees: an overview of likelihood-based linkage methods. Hum Hered. 2007;64(3):192–202.
9 Risch N. Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet. 1984;36(2):363–86.
10 Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J. Effects of misspecifying genetic parameters in lod score analysis. Biometrics. 1986;42(2):393–9.
11 Kraft P, Thomas DC. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet. 2000;66(3):1119–31.
12 Elston RC. Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. Am J Med Genet. 1989;34(4):487–8.
13 Ginsburg E, Malkin I, Elston RC. Sampling correction in linkage analysis. Genet Epidemiol. 2004;27(2):87–96.
14 Malkin I, Elston RC. Response to letter by Veronica J. Vieland and Susan E. Hodge. Genet Epidemiol. 2005;28(3):286–7.
15 Brugger M, Rospleszcz S, Strauch K. Estimation of trait-model parameters in a MOD score linkage analysis. Hum Hered. 2016;82(3–4):103–39.
16 Strauch K. Parametric linkage analysis with automatic optimization of the disease model parameters. Am J Hum Genet. 2003;73(Suppl 1):A2624.
17 Dietter J, Mattheisen M, Fürst R, Rüschendorf F, Wienker TF, Strauch K. Linkage analysis using sex-specific recombination fractions with GENEHUNTER-MODSCORE. Bioinformatics. 2007;23(1):64–70.
18 Mattheisen M, Dietter J, Knapp M, Baur MP, Strauch K. Inferential testing for linkage with GENEHUNTER-MODSCORE: the impact of the pedigree structure on the null distribution of multipoint MOD scores. Genet Epidemiol. 2008;32(1):73–83.
19 Brugger M, Strauch K. Fast linkage analysis with MOD scores using algebraic calculation. Hum Hered. 2014;78(3–4):179–94.
20 Künzel T, Strauch K. Parameter estimation and quantitative parametric linkage analysis with GENEHUNTER-QMOD. Hum Hered. 2012;73(4):208–19.
21 Lewis CM, Knight J. Introduction to genetic association studies. Cold Spring Harb Protoc. 2012;2012(3):297–306.
22 Hodge SE, Boehnke M, Spence MA. Loss of information due to ambiguous haplotyping of SNPs. Nat Genet. 1999;21(4):360–1.
23 Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet. 1987;51(3):227–33.
24 Terwilliger JD, Ott J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered. 1992;42(6):337–46.
25 Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993;52(3):506–16.
26 Curtis D. Use of siblings as controls in case-control association studies. Ann Hum Genet. 1997;61(Pt 4):319–33. Erratum in: Ann Hum Genet 1998 Jan;62(Pt 1):89.
27 Martin ER, Kaplan NL, Weir BS. Tests for linkage and association in nuclear families. Am J Hum Genet. 1997;61(2):439–48.
28 Boehnke M, Langefeld CD. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet. 1998;62(4):950–61.
29 Lazzeroni LC, Lange K. A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered. 1998;48(2):67–81.
30 Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet. 1998;62(2):450–8.
31 Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. Am J Hum Genet. 1999;64(3):861–70.
32 Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet. 2000;67(1):146–54.
33 Wicks J. Exploiting excess sharing: a more powerful test of linkage for affected sib pairs than the transmission/disequilibrium test. Am J Hum Genet. 2000;66(6):2005–8.
34 Wicks J, Wilson SR. Evaluating linkage and linkage disequilibrium: use of excess sharing and transmission disequilibrium methods in affected sib pairs. Ann Hum Genet. 2000;64(Pt 5):419–32.
35 Lazzeroni LC. Allele sharing and allelic association I: sib pair tests with increased power. Genet Epidemiol. 2002;22(4):328–44.
36 Xiong M, Jin L. Combined linkage and linkage disequilibrium mapping for genome screens. Genet Epidemiol. 2000;19(3):211–34.
37 Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. Genet Epidemiol. 2000;19(Suppl 1):S36–42.
38 Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered. 2000;50(4):211–23.
39 Allen-Brady K, Wong J, Camp NJ. PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. BMC Bioinf. 2006;7:209.
40 Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. Am J Hum Genet. 2000;66(1):279–92.

41 Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet. 1999;65(4):1170–7.
42 Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered. 2008;66(2):87–98.
43 Jorde LB. Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet. 1995;56(1):11–4.
44 Clerget-Darpoux F. Bias of the estimated recombination fraction and lod score due to an association between disease gene and a marker gene. Ann Hum Genet. 1982;46:363–372.
45 Huang Q, Shete S, Amos CI. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. Am J Hum Genet. 2004;75(6):1106–12.
46 Boyles AL, Scott WK, Martin ER, Schmidt S, Li YJ, Ashley-Koch A, et al. Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. Hum Hered. 2005;59(4):220–7.
47 Levinson DF, Holmans P. The effect of linkage disequilibrium on linkage analysis of incomplete pedigrees. BMC Genet. 2005;6(Suppl 1):S6.
48 Kim Y, Duggal P, Gillanders EM, Kim H, Bailey-Wilson JE. Examining the effect of linkage disequilibrium between markers on the Type I error rate and power of non-parametric multipoint linkage analysis of two-generation and multigenerational pedigrees in the presence of missing genotype data. Genet Epidemiol. 2008;32(1):41–51.
49 MacLean CJ, Morton NE, Yee S. Combined analysis of genetic segregation and linkage under an oligogenic model. Comput Biomed Res. 1984;17(5):471–80.
50 Clerget-Darpoux F, Babron MC, Prum B, Lathrop GM, Deschamps I, Hors J. A new method to test genetic models in HLA associated diseases: the MASC method. Ann Hum Genet. 1988;52(3):247–58.
51 Tienari PJ, Wikström J, Sajantila A, Palo J, Peltonen L. Genetic susceptibility to multiple sclerosis linked to myelin basic protein gene. Lancet. 1992;340(8826):987–91.
52 Fan R, Xiong M. Combined high resolution linkage and association mapping of quantitative trait loci. Eur J Hum Genet. 2003;11(2):125–37.
53 Jung J, Fan R, Jin L. Combined linkage and association mapping of quantitative trait loci by multiple markers. Genetics. 2005;170(2):881–98.
54 Hasstedt SJ. Version 7.1 Pedigree Analysis Package. Salt Lake City: Department of Human Genetics University of Utah; 2009.
55 Hasstedt SJ, Thomas A. Detecting pleiotropy and epistasis using variance components linkage analysis in jPAP. Hum Hered. 2011;72(4):258–63.
56 Lathrop GM, Lalouel JM. Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet. 1984;36(2):460–5.
57 Lathrop GM, Lalouel JM, Julier C, Ott J. Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci U S A. 1984;81(11):3443–6.
58 Lathrop GM, Lalouel JM, White RL. Construction of human linkage maps: likelihood calculations for multilocus linkage analysis. Genet Epidemiol. 1986;3(1):39–52.
59 Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, et al. MENDEL version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. Am J Hum Genet. 2001;69(Suppl l):504.
60 Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM. Mendel: the Swiss army knife of genetic analysis programs. Bioinformatics. 2013;29(12):1568–70.
61 Li M, Boehnke M, Abecasis GR. Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. Am J Hum Genet. 2005;76(6):934–49.
62 Li M, Boehnke M, Abecasis GR. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am J Hum Genet. 2006;78(5):778–92.
63 Hiekkalinna T, Schäffer AA, Lambert B, Norrgrann P, Göring HH, Terwilliger JD. PSEUDOMARKER: a powerful program for joint linkage and/or linkage disequilibrium analysis on mixtures of singletons and related individuals. Hum Hered. 2011;71(4):256–66.
64 Gertz EM, Hiekkalinna T, Digabel SL, Audet C, Terwilliger JD, Schäffer AA. PSEUDOMARKER 2.0: efficient computation of likelihoods using NOMAD. BMC Bioinf. 2014;15:47.
65 Lou XY, Casella G, Todhunter RJ, Yang MCK, Wu R. A general statistical framework for unifying interval and linkage disequilibrium mapping: toward high-resolution mapping of quantitative traits. J Am Stat Assoc. 2005;100(469):158–71.
66 Cantor RM, Chen GK, Pajukanta P, Lange K. Association testing in a linked region using large pedigrees. Am J Hum Genet. 2005;76(3):538–42.
67 Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP. Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. Am J Hum Genet. 2000;66(6):1945–57.
68 Horsthemke B. In brief: genomic imprinting and imprinting diseases. J Pathol. 2014;232(5):485–7.
69 Brugger M, Knapp M, Strauch K. Properties and evaluation of the MOBIT - a novel linkage-based test statistic and quantification method for imprinting. Stat Appl Genet Mol Biol. 2019;18(4).
70 Vieland VJ, Hodge SE. Ascertainment bias in linkage analysis: comments on Ginsburg et al. Genet Epidemiol. 2005;28(3):283–7; author reply 286—7.
71 Becker T, Herold C. Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power. Eur J Hum Genet. 2009;17(8):1043–9.
72 Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol. 2002;23(3):221–33.
73 Balliu B, Houwing-Duistermaat JJ, Böhringer S. Powerful testing via hierarchical linkage disequilibrium in haplotype association studies. Biom J. 2019;61(3):747–68.
74 Ceppellini R, Siniscalco M, Smith CA. The estimation of gene frequencies in a random-mating population. Ann Hum Genet. 1955;20(2):97–115.
75 Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the E-M algorithm. J R Stat Soc Ser B. 1977;39(1):1–22.
76 Risch N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet. 1990;46(2):229–41.
77 Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A. 1987;84(8):2363–7.
78 Abecasis GR, Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet. 2005;77(5):754–67.
79 Boehnke M. Allele frequency estimation from data on relatives. Am J Hum Genet. 1991;48(1):22–5.
80 Göring HH, Terwilliger JD. Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. Am J Hum Genet. 2000;66(4):1298–309.
81 Rohde K, Fuerst R. Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. Hum Mutat. 2001;17(4):289–95.
82 Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002;30(1):97–101.
83 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet. 1996;58(6):1347–63.
84 Gao G, Allison DB, Hoeschele I. Haplotyping methods for pedigrees. Hum Hered. 2009;67(4):248–66.
85 Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet. 1996;58(6):1323–37.

86 Neugebauer M. Mathematische Methoden und Algorithmen zur Analyse genetischer Polymorphismen in Stammbäumen. Inaugural dissertation. University of Bonn; 1989.
87 Powell MJD. A direct search optimization method that models the objective and constraint functions by linear interpolation. In: Gomez S, Hennart JP, editors. Advances in optimization and numerical analysis. Dordrecht: Kluwer Academic; 1994. p. 51—67.
88 Powell MJD. Direct search algorithms for optimization calculations. Acta Numer. 1998;7:287–336.
89 Johnson SG. The NLopt nonlinear-optimization package; 2020. http://github.com/stevengj/nlopt.
90 Altieri D, Tubaldi E, De Angelis M, Patelli E, Dall'Asta A. Reliability-based optimal design of nonlinear viscous dampers for the seismic protection of structural systems. Bull Earthquake Eng. 2018;16(2):963–82.
91 Cho K, Dupuis J. Handling linkage disequilibrium in qualitative trait linkage analysis using dense SNPs: a two-step strategy. BMC Genet. 2009;10:44.
92 Abecasis GR, Cookson WO. GOLD—graphical overview of linkage disequilibrium. Bioinformatics. 2000;16(2):182–3.
93 Ott J. Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci U S A. 1989;86(11):4175–8.
94 Weeks DE, Lehner T, Squires-Wheeler E, Kaufmann C, Ott J. Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis. Genet Epidemiol. 1990;7(4):237–43.
95 Schäffer AA, Lemire M, Ott J, Lathrop GM, Weeks DE. Coordinated conditional simulation with SLINK and SUP of many markers linked or associated to a trait in large pedigrees. Hum Hered. 2011;71(2):126–34.
96 Shete S, Zhou X. Parametric approach to genomic imprinting analysis with applications to Angelman's syndrome. Hum Hered. 2005;59(1):26–33.
97 Lemire M. SUP: an extension to SLINK to allow a larger number of marker loci to be simulated in pedigrees conditional on trait values. BMC Genet. 2006;7:40.
98 Llach J, Carballal S, Moreira L. Familial pancreatic cancer: Current perspectives. Cancer Manag Res. 2020;12:743–58.
99 Bartsch DK, Gress TM, Langer P. Familial pancreatic cancer—current knowledge. Nat Rev Gastroenterol Hepatol. 2012;9(8):445–53.
100 Bartsch DK, Sina-Frey M, Ziegler A, Hahn SA, Przypadlo E, Kress R, et al. Update of familial pancreatic cancer in Germany. Pancreatology. 2001;1(5):510–6.
101 Bartsch DK, Matthäi E, Mintziras I, Bauer C, Figiel J, Sina-Boemers M, et al. The German national case collection for familial pancreatic cancer (FaPaCa)—knowledge gained in 20 years. Dtsch Arztebl Int. 2021;118:163–8.
102 Lehman B, Matthäi E, Gercke N, Denzer UW, Figiel J, Hess T, et al. Characteristics of familial pancreatic cancer families with additional colorectal carcinoma. Fam Cancer. 2023;22(3):323–30.
103 Seeber A, Zimmer K, Kocher F, Puccini A, Xiu J, Nabhan C, et al. Molecular characteristics of BRCA1/2 and PALB2 mutations in pancreatic ductal adenocarcinoma. ESMO Open. 2020;5(6):e000942.
104 Goldstein JI, Crenshaw A, Carey J, Grant GB, Maguire J, Fromer M, et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinformatics. 2012;28(19):2543–5.
105 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
106 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–73.
107 Dimitromanolakis A, Paterson AD, Sun L. Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data via TRUFFLE. Am J Hum Genet. 2019;105(1):78–88.
108 Hiekkalinna T, Göring HH, Terwilliger JD. On the validity of the likelihood ratio test and consistency of resulting parameter estimates in joint linkage and linkage disequilibrium analysis under improperly specified parametric models. Ann Hum Genet. 2012;76(1):63–73.
109 Nato AQ, Buyske S, Matise TC. The Rutgers map: a third-generation combined linkage-physical map of the human genome. Manuscript in preparation. Available from: http://compgen.rutgers.edu/download_maps.shtml.
110 Oikkonen J, Kuusi T, Peltonen P, Raijas P, Ukkola-Vuoti L, Karma K, et al. Creative activities in music: a genome-wide linkage analysis. PLoS One. 2016;11(2):e0148679.
111 Handel-Fernandez ME, Nassiri M, Arana M, Perez MM, Fresno M, Nadji M, et al. Mapping of genetic deletions on the long arm of chromosome 22 in human pancreatic adenocarcinomas. Anticancer Res. 2000;20(6B):4451–6.
112 Wild A, Langer P, Celik I, Chaloupka B, Bartsch DK. Chromosome 22q in pancreatic endocrine tumors: identification of a homozygous deletion and potential prognostic associations of allelic deletions. Eur J Endocrinol. 2002;147(4):507–13.
113 Li YY, Mukaida N. Pathophysiological roles of Pim-3 kinase in pancreatic cancer development and progression. World J Gastroenterol. 2014;20(28):9392–404.
114 Franke A, Balschun T, Sina C, Ellinghaus D, Häsler R, Mayr G, et al. Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). Nat Genet. 2010;42(4):292–4.
115 Everhov ÅH, Erichsen R, Sachs MC, Pedersen L, Halfvarson J, Askling J, et al. Inflammatory bowel disease and pancreatic cancer: a Scandinavian register-based cohort study 1969-2017. Aliment Pharmacol Ther. 2020;52(1):143–54.
116 Lowenfels AB, Maisonneuve P, DiMagno EP, Elitsur Y, Gates LK Jr, Perrault J, et al. Hereditary pancreatitis and the risk of pancreatic cancer. International Hereditary Pancreatitis Study Group. J Natl Cancer Inst. 1997;89(6):442–6.
117 Paterson AD, Naimark DM, Petronis A. The analysis of parental origin of alleles may detect susceptibility loci for complex disorders. Hum Hered. 1999;49(4):197–204.