Aus dem
Lehrstuhl für Genetische Epidemiologie, IBE
Institut der Universität München
Leiter: Prof. Dr. Ulrich Mansmann

Methoden der Kopplungs- und Assoziationsanalyse in Familien

Dissertation
zum Erwerb des Doktorgrades der Humanbiologie
an der Medizinischen Fakultät
der Ludwig-Maximilians-Universität zu München

vorgelegt von
Markus Fridolin Brugger

aus
Donaueschingen

Jahr
2025

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	Prof. Dr. Konstantin Strauch
Mitberichterstatter:	Prof. Dr. Ortrud Steinlein
	Prof. Dr. Annika Hoyer
	PD Dr. Larissa Schwarzkopf
Mitbetreuung durch den promovierten Mitarbeiter:	
Dekan:	Prof. Dr. med. Thomas Gudermann
Tag der mündlichen Prüfung:	07.02.2025

Affidavit

Eidesstattliche Versicherung

Brugger, Markus Fridolin

Name, Vorname

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel:

Methoden der Kopplungs- und Assoziationsanalyse in Familien

……

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Seifhennersdorf, 24.2.2025 Markus Fridolin Brugger

Ort, Datum Unterschrift Doktorandin bzw. Doktorand

Inhaltsverzeichnis

Abkürzungsverzeichnis

cM	centiMorgan
DNA	engl. *desoxyribonucleic acid*, Desoxyribonukleinsäure
FaPaCa	Nationale Fallsammlung für familiäres Pankreaskarzinom
GHM	GENEHUNTER-MODSCORE
GWAS	engl. *genome-wide association study*, genomweite Assoziationsstudie
LD	engl. *linkage disequilibrium*, Kopplungsungleichgewicht
LE	engl. *linkage equilibrium*, Kopplungsgleichgewicht
SNV	engl. *single nucleotide variant*, Einzelnukleotidvariante
WGS	engl. *whole-genome sequencing*, vollständige Genomsequenzierung

Publikationsliste

Brugger M, Lutz M, Müller-Nurasyid M, Lichtner P, Slater EP, Matthäi E, Bartsch DK, Strauch K. Joint linkage and association analysis with GENEHUNTER-MODSCORE with an application to familial pancreatic cancer. Hum Hered. 2024;89(1):8—31.

Brugger M. Untersuchungen zum Einfluss des Bibers (*Castor fiber*) auf den Fischotter (*Lutra lutra*) anhand von Aktivitätsdichteschätzungen in der Oberlausitz. Arbeitskreis Biberschutz im NABU Landesverband Sachsen-Anhalt e. V., Mitteilungen des Arbeitskreises Biberschutz. 2022;25—34.

Landmann E, **Brugger M**, Blank V, Wudy SA, Hartmann M, Strauch K, Rudloff S. Adrenal steroid metabolism and blood pressure in 5- to 7-year-old children born preterm as compared to peers born at term. Front Pediatr. 2021;9:754989.

Brugger M, Jährig M, Peper J, Nowak C, Cocchiararo B, Ansorge H. Influence of Eurasian beaver (*Castor fiber*) on Eurasian otter (*Lutra lutra*) evaluated by activity density estimates in anthropogenic habitats in eastern Germany. IUCN Otter Spec Group Bull. 2020;37(2):98—119.

Burchert A, Bug G, Fritz LV, Finke J, Stelljes M, Röllig C, Wollmer E, Wäsch R, Bornhäuser M, Berg T, Lang F, Ehninger G, Serve H, Zeiser R, Wagner EM, Kröger N, Wolschke C, Schleuning M, Götze KS, Schmid C, Crysandt M, Eßeling E, Wolf D, Wang Y, Böhm A, Thiede C, Haferlach T, Michel C, Bethge W, Wündisch T, Brandts C, Harnisch S, Wittenberg M, Hoeffkes HG, Rospleszcz S, Burchardt A, Neubauer A, **Brugger M**, Strauch K, Schade-Brittinger C, Metzelder SK. Sorafenib maintenance after allogeneic hematopoietic stem cell transplantation for acute myeloid leukemia with *FLT3*-internal tandem duplication mutation (SORMAIN). J Clin Oncol. 2020;38(26):2993—3002.

Schubert R, **Brugger M**, Kühnel S, Hohlfeld H, Heidger CM. Analyses of sexual reproductive traits in *Dactylorhiza majalis*: a case study from East Germany. Biologia. 2020;75:507—521.

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):20180025.

Brugger M. Habitatansprüche des Fischotters im anthropogenen Lebensraum der Oberlausitz. Beiträge zur Jagd- und Wildforschung. 2018;43:343—364.

Jegan N, **Brugger M**, Viniol A, Strauch K, Barth J, Baum E, Leonhardt C, Becker A. Psychological risk and protective factors for disability in chronic low back pain - a longitudinal analysis in primary care. BMC Musculoskelet Disord. 2017;18(1):114.

Brugger M, Rospleszcz S, Strauch K. Estimation of trait-model parameters in a MOD score linkage analysis. Hum Hered. 2016;82(3-4):103—139.

Teymoortash A, Pfestroff A, Wittig A, Franke N, Hoch S, Harnisch S, Schade-Brittinger C, Hoeffken H, Engenhart-Cabillic R, **Brugger M**, Strauch K. Safety and efficacy of botulinum toxin to preserve gland function after radiotherapy in patients with head and neck cancer: a prospective, randomized, placebo-controlled, double-blinded phase I clinical trial. PLoS ONE. 2016;11(3):e0151316.

Viniol A, Jegan N, **Brugger M**, Leonhardt C, Barth J, Baum E, Becker A, Strauch K. Even worse – risk factors and protective factors for transition from chronic localized low back pain to chronic widespread pain in general practice: a cohort study. Spine. 2015;40(15):E890—E899.

Brugger M, Strauch K. Fast linkage analysis with MOD scores using algebraic calculation. Hum Hered. 2014;78(3-4):179—194.

Petersen AK, Zeilinger S, Kastenmüller G, Römisch-Margl W, **Brugger M**, Peters A, Meisinger C, Strauch K, Hengstenberg C, Pagel P, Huber F, Mohney RP, Gallert H, Illig T, Adamski J, Waldenberger M, Gieger C, Suhre K. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. Hum Mol Genet. 2014;23(2):534—545.

Huke V, Rudloff S, **Brugger M**, Strauch K, Berthold LD, Landmann E. Prematurity is not associated with intra-abdominal adiposity in 5- to 7-year-old children. J Pediatr. 2013;163(5):1301—1306.

Viniol A, Jegan N, Hirsch O, Leonhardt C, **Brugger M**, Strauch K, Barth J, Baum E, Becker A. Chronic low back pain patient groups in primary care – a cross sectional cluster analysis. BMC Musculoskelet Disord. 2013;14:294.

Viniol A, Jegan N, Leonhardt C, **Brugger M**, Strauch K, Barth J, Baum E, Becker A. Differences between patients with chronic widespread pain and local chronic low back pain in primary care – a comparative cross-sectional analysis. BMC Musculoskelet Disord. 2013;14:351.

Faryna M, Konermann C, Aulmann S, Bermejo JL, **Brugger M**, Diederichs S, Rom J, Weichenhan D, Claus R, Rehli M, Schirmacher P, Sinn HP, Plass C, Gerhauser C. Genome-wide methylation screen in low-grade breast cancer identifies novel epigenetically altered genes as potential biomarkers for tumor diagnosis. FASEB J. 2012;26(12):4937—4950.

Viniol A, Jegan N, Leonhardt C, Strauch K, **Brugger M**, Barth J, Baum E, Becker A. Study protocol: Transition from localized low back pain to chronic widespread pain

in general practice: identification of risk factors, preventive factors and key elements for treatment – a cohort study. BMC Musculoskelet Disord. 2012;13:77.

Heinzmann A, **Brugger M**, Bierbaum S, Mailaparambil B, Kopp MV, Strauch K. Joint influences of acidic-mammalian-chitinase with interleukin-4 and toll-like receptor-10 with interleukin-13 in the genetics of asthma. Pediatr Allergy Immunol. 2010;21(4 Pt 2):e679—e686.

Heinzmann A, **Brugger M**, Engels C, Prömpeler H, Superti-Furga A, Strauch K, Krueger M. Risk factors of neonatal respiratory distress following vaginal delivery and caesarean section in the German population. Acta Paediatr. 2009;98(1):25—30.

Konferenzbeiträge (Poster)

Brugger M, Lutz M, Müller-Nurasyid M, Lichtner P, Slater EP, Matthäi E, Bartsch DK, Strauch K. Joint linkage and association analysis with GENEHUNTER-MODSCORE with an application to the German national case collection of familial pancreatic cancer (FaPaCa). 18. Jahrestagung der Deutschen Gesellschaft für Epidemiologie (DGEpi), Würzburg, 2023

Brugger M, Lutz M, Strauch K. Joint Linkage and Association Analysis with GENEHUNTER-MODSCORE. 31st annual meeting of the International Genetic Epidemiology Society (IGES), Paris, 2022

Brugger M, Strauch K. Fast linkage analysis with MOD scores using algebraic calculation. 23rd annual meeting of the International Genetic Epidemiology Society (IGES), Wien, 2014

Brugger M, Lemire M, Knapp M, Strauch K. Confounding between genomic imprinting and sex-specific recombination frequencies: evaluation of properties of the MOD score-based imprinting test statistic MOBIT in a linkage simulation study. 20th annual meeting of the International Genetic Epidemiology Society (IGES), Heidelberg, 2012

1. Beiträge zu den Veröffentlichungen

1.1 Beitrag zu Paper I: Brugger und Strauch, 2014

Ich habe die in diesem Paper beschriebene algorithmische Optimierung einer kopplungsanalytischen Methode nach einer Idee von Prof. Strauch eigenständig entwickelt und in das Softwareprogramm GENEHUNTER-MODSCORE implementiert. Die zur Evaluierung der neuen Methode nötige Simulationsstudie habe ich selbst entworfen und durchgeführt. Des Weiteren war ich verantwortlich für die Konzeption und die Erstellung des Manuskripts.

1.2 Beitrag zu Paper II: Brugger et al., 2024

Ich habe die in diesem Paper beschriebene neue Version von GENEHUNTER-MODSCORE, die eine gemeinsame Kopplungs- und Assoziationsanalyse mit Familien und unverwandten Individuen erlaubt, eigenständig entwickelt und implementiert. Die zur Evaluierung der neuen Methode nötige Simulationsstudie habe ich selbst entworfen und durchgeführt. Die MOD-Score-Analyse der Daten aus der Nationalen Fallsammlung für familiäres Pankreaskarzinom (*German National Case Collection for Familial Pancreatic Cancer* (FaPaCa)) als Anwendungsbeispiel, für deren Bereitstellung und erste Qualitätskontrollen weitere Koautoren des Papers (Manuel Lutz, Martina Müller-Nurasyid, Elvira Matthäi, Emily P. Slater und Detlef K. Bartsch) verantwortlich waren, habe ich ebenfalls selbständig durchgeführt. Des Weiteren war ich verantwortlich für die Konzeption und die Erstellung des Manuskripts. Die Idee für eine gemeinsame Kopplungs- und Assoziationsanalyse mit Familien und unverwandten Individuen im Rahmen einer MOD-Score-Analyse geht auf Prof. Strauch zurück.

1.3 Beitrag zu Paper III (Anhang A): Brugger et al., 2016

Die in diesem Paper veröffentlichte Simulationsstudie zur Schätzbarkeit von Krankheitsmodellparametern im Rahmen einer MOD-Score-Analyse basiert auf einer Idee Prof. Strauchs. Dabei habe ich, basierend auf Vorarbeiten bezüglich Design der Studie und Datenanalyse durch die Koautorin Susanne Rospleszcz, die Studie eigenständig durchgeführt. Ich habe die Zusammenstellung der bisherigen Erkenntnisse aus der Literatur zur Schätzbarkeit von Krankheitsmodellparametern im Rahmen von Kopplungsanalysen eigenständig unternommen und war verantwortlich für die Konzeption und die Erstellung des Manuskripts.

1.4 Beitrag zu Paper IV (Anhang B): Brugger et al., 2019

Die in diesem Paper veröffentlichte Untersuchung des Confoundings zwischen geschlechtsspezifischen Rekombinationsfrequenzen und genomischem Imprinting bei der Kopplungsanalyse inklusive Simulationsstudie habe ich, basierend auf einer Idee Prof. Strauchs, eigenständig konzipiert und durchgeführt. Bei der formalen Vorstellung der Teststatistik MOBIT als Test auf genomisches Imprinting in Anwesenheit von Kopplung im Rahmen einer MOD-Score-Analyse war ich maßgeblich gemeinsam mit dem Koautor Michael Knapp und Prof. Strauch beteiligt. Ich war verantwortlich für die Konzeption und die Erstellung des Manuskripts und habe die im Manuskript erwähnte, neue Simulationsprozedur für den MOBIT in GENEHUNTER-MODSCORE selbständig implementiert.

2. Einleitung zu den Publikationen

2.1 Übersicht über die in dieser Arbeit verwendeten Publikationen

Die in dieser Arbeit als Paper I bezeichnete Publikation (Brugger und Strauch, 2014) befasst sich mit einer algorithmischen Optimierung im Zusammenhang mit einer kopplungsanalytischen Methode und deren Implementation in das Softwarepaket GENEHUNTER-MODSCORE (Strauch, 2003, Dietter et al., 2007, Mattheisen et al., 2008, Brugger und Strauch, 2014, Brugger et al., 2024). Basierend auf den algorithmischen Optimierungen in Paper I entstanden in der Folge zwei weitere Publikationen (Paper III (Brugger et al., 2016) und Paper IV (Brugger et al., 2019), siehe Anhänge A und B), die sich mit der Schätzbarkeit von Krankheitsmodellparametern (Paper III) und der Testung auf genomisches Imprinting (Paper IV) im Rahmen der kopplungsanalytischen Methode, wie sie in GENEHUNTER-MODSCORE implementiert ist, befassen. In Paper II (Brugger et al., 2024) geht es schließlich um die gemeinsame Kopplungs- und Assoziationsanalyse in Familien unter Hinzunahme unverwandter Individuen und deren Implementation in GENEHUNTER-MODSCORE. Im Folgenden wird es eine kurze Einleitung zum Thema Krankheitsgenkartierung geben. Anschließend werden die Paper I und II sowie ergänzend auch die Paper III und IV kurz zusammenhängend dargestellt.

2.2 Allgemeiner Teil: Die Krankheitsgenkartierung

Träger der Erbinformation ist die DNA (engl. *desoxyribonucleic acid*, Desoxyribonukleinsäure), die beim Menschen als 22 Körperchromosomen (Autosomen) und 1 Paar Geschlechtschromosomen (Gonosomen) im Zellkern vorliegen. In einem somatischen Zellkern liegt jedes Chromosom in zweifacher Kopienzahl vor, der Mensch ist also diploid (2n = 46). Dabei stammt je ein Chromosom eines homologen Chromosomenpaares von der Mutter (maternale Herkunft) und eines vom Vater (paternale Herkunft). Die Definition des Begriffes „Gen“ ist nach wie vor einem stetigen Wandel unterworfen (siehe Portin und Wilkins, 2017), die für die vorliegende Arbeit hinreichende Definition eines Gens kann folgendermaßen formuliert werden: Ein Gen ist eine spezifische Region auf der DNA, welche für ein Enzym oder Protein codiert, wobei die Position eines Gens auf einem Chromosom als Locus und Variationen der DNA-Sequenz an diesem Locus als Allele bezeichnet werden (Weiss, 1993). Das Paar von Allelen an einem Locus eines Individuums bezeichnet man als Genotyp, Allele an verschiedenen Loci, die vom selben Elternteil stammen, als Haplotyp (Ott, 1999). Die Kartierung eines potenziell krankheitsursächlichen Gens für eine Erbkrankheit, deren molekulargenetische Grundlage unbekannt ist, erfolgt in der Humangenetik unter Verwendung des

Ansatzes der Positionsklonierung (Botstein und Risch, 2003). Hierbei wird der Krankheitslocus zuerst kartiert, bevor die Genfunktion untersucht wird, was der Methode den Beinamen „reverse Genetik“ einbrachte (Ruddle, 1984). Bei der Positionsklonierung wird in einem ersten Schritt eine Kopplungsanalyse durchgeführt, um die chromosomale Position des mutmaßlichen Krankheitsgenortes einzugrenzen (Grobkartierung), gefolgt von einer Feinkartierung mittels einer Assoziationsanalyse (Botstein und Risch, 2003). Eine Kopplungsanalyse untersucht die gemeinsame Vererbung eines Merkmals, wie z. B. dem Krankheitsstatus gesund/krank, mit einem genetischen Marker in Familien. Innerhalb eines Familienstammbaumes werden sogenannte Founder von Nonfoundern unterschieden. Die Eltern der Founder sind im Stammbaum nicht verfügbar und begründen diesen daher, sie gelten als zufällig aus der Gesamtbevölkerung gezogen, sodass man verschiedene Stammbäume als unabhängig voneinander betrachtet. Nonfounder hingegen sind Nachkommen, deren Eltern im Stammbaum verfügbar sind. Als genetischer Marker kommt im Prinzip jeder genetische Polymorphismus infrage, also z. B. Einzelnukleotidvarianten (engl. *single nucleotide variants*, SNVs) oder auch hochpolymorphe Mikrosatelliten, die aus mehrfachen, kurzen Tandemwiederholungen derselben DNA-Sequenz bestehen. Liegen Marker- und Krankheitsgenort nahe genug auf einem Chromosom beieinander, so werden die beiden dazugehörigen, sich auf einem elterlichen Chromosom befindlichen Allele öfter gemeinsam an die Nachkommen vererbt, als dies unter Zufallsbedingungen der Fall wäre, und es liegt Kopplung zwischen Marker- und Krankheitsgenort vor. Bei zunehmender Distanz zwischen Marker- und Krankheitsgenort werden die beiden entsprechenden Merkmale zunehmend öfter voneinander während der Meiose I (Reduktionsteilung) durch Crossover-Ereignisse der homologen, elterlichen Chromosomen getrennt und neu rekombiniert. Das entsprechende Maß für die Kopplung zweier Genorte ist die Rekombinationsfrequenz θ, die zwischen 0 (vollständiger Kopplung) und 0,5 (keiner Kopplung) rangiert. Die genetische Kopplung zwischen zwei Genorten stellt demzufolge eine Abweichung vom 3. Mendelschen Gesetz, der Unabhängigkeitsregel, dar. Die räumliche Auflösung einer Kopplungsanalyse liegt typischerweise zwischen 2,5 und 10 centiMorgan (cM) (Fan und Jung, 2003), sodass in der Folge häufig eine Assoziationsanalyse zwecks Feinkartierung nachgeschaltet wird. An dieser Stelle sei darauf hingewiesen, dass im Rahmen der Krankheitsgenkartierung heutzutage oft direkt eine genomweite Assoziationsstudie (engl. *genome-wide association study*, GWAS) ohne vorgeschaltete Kopplungsanalyse gemacht wird. Eine solche GWAS ist im Falle von komplexen Krankheiten, die zu einem erheblichen Anteil von häufigen genetischen Varianten mit moderatem Effekt mitverursacht werden, in der Regel trennschärfer als die Kopplungsanalyse und kommt auch gänzlich ohne Familiendaten aus (Ott et al., 2015). Hingegen ist im Falle von Krankheiten, die eine familiäre Häufung aufweisen und meist durch seltene Varianten mit großem Effekt verursacht werden,

die Kopplungsanalyse der GWAS überlegen (Ott et al., 2015). Dabei ist die Nutzbarkeit einer Kopplungsanalyse zur Kartierung von seltenen krankheitsursächlichen Varianten in den letzten Jahren durch die Verfügbarkeit von Daten aus vollständiger Genomsequenzierung (engl. *whole-genome sequencing*, WGS) stark angestiegen. Die Kopplungsanalyse kann dabei auch mit bioinformatischen Methoden zur Auswahl und Prädiktion potenziell schädlicher Varianten anhand von WGS-Daten in Familien kombiniert werden (Ott et al., 2015).

Eine Assoziationsanalyse untersucht die Korrelation von Markerallelen mit dem Krankheitszustand von unverwandten Betroffenen und Gesunden aus der Bevölkerung oder auch in Familien (Terwilliger, 1995). Im Falle keiner Korrelation befinden sich die Allele am Markergenort mit denen des Krankheitsgenorts im Kopplungsgleichgewicht (engl. *linkage equilibrium*, LE), wodurch die Verteilung der Allele am Markergenort bei den Betroffenen und Gesunden gleich ist. Eine Korrelation von Markerallelen mit dem Krankheitsgenort liegt vor, wenn sich deren Verteilung zwischen den Betroffenen und Gesunden unterscheidet, was dann Kopplungsungleichgewicht (engl. *linkage disequilibrium*, LD) genannt wird.

Des Weiteren kann es auch vorteilhaft sein, beide Kartierungsstrategien, also Kopplung und Assoziation, direkt miteinander in einer gemeinsamen Kopplungs- und Assoziationsanalyse zu verbinden, was insbesondere dann zu einer erhöhten Trennschärfe der statistischen Analyse führen kann, wenn sowohl größere Familien als auch Unverwandte für die Analyse zur Verfügung stehen (Göring und Terwilliger, 2000).

Meine beiden im Folgenden dargestellten Hauptarbeiten befassen sich mit algorithmischen Optimierungen im Rahmen einer speziellen kopplungsanalytischen Methode (Paper I) und der Entwicklung eines gemeinsamen Kopplungs- und Assoziationsverfahrens für Datensätze aus Familien und Unverwandten (Paper II).

2.3 Paper I: Algorithmische Optimierung einer kopplungsanalytischen Methode

Die Kopplungsanalyse ist eine bewährte Methode, Krankheiten, die einer familiären Häufung unterliegen, im Genom zu kartieren (Ott et al., 2015). Kopplungsanalysen werden traditionell in „parametrische“ (modellbasierte) oder „nicht-parametrische“ (modellfreie) Analysen unterteilt. In der parametrischen Kopplungsanalyse, die auch unter dem Namen LOD-Score-Analyse bekannt ist (Morton, 1955), werden bestimmte Krankheitsmodellparameter für die Vererbung der Krankheit angenommen, wohingegen nicht-parametrische Methoden ohne eine solche Annahme auskommen. Es konnte jedoch gezeigt werden, dass bestimmte parametrische und nicht-parametrische Verfahren statistisch äquivalent zueinander sind (Knapp et al., 1994, Strauch,

2007). Die im Folgenden gemachten Annahmen bezüglich des Krankheitsmodells und der dafür nötigen Krankheitsmodellparameter gelten für alle in dieser Arbeit vorgestellten Paper. Im Falle eines diallelischen autosomalen Locus, der eine dichotome Krankheit verursacht, sind die Krankheitsmodellparameter folgende: die Krankheitsallelfrequenz p_m („m“ steht für Mutation, mit Wildtypallelfrequenz $p_+ = 1 - p_m$), die drei Penetranzen f_0, f_1, f_2, wobei f_i die Wahrscheinlichkeit beschreibt, mit der ein Individuum mit i Kopien des Krankheitsallels von der Krankheit betroffen ist, und die Rekombinationsfrequenz θ zwischen Marker- und Krankheitslocus bzw. die genetische Position x des Krankheitslocus im Falle einer Analyse mit mehr als einem genetischen Marker (Multimarkeranalyse, Multipointanalyse). Hängt die Expression eines krankheitsursächlichen Gens von dessen elterlicher Herkunft ab, so liegt genomisches Imprinting vor (Hall, 1990), welches durch die Aufspaltung der Heterozygotenpenetranz f_1 in zwei Penetranzen $f_{1,mat}$ (maternale Herkunft) und $f_{1,pat}$ (paternale Herkunft des Krankheitsallels) im Rahmen einer LOD-Score-Analyse berücksichtigt werden kann (Strauch et al., 2000). Die zur Berechnung des LOD-Scores als Kopplungstest notwendigen Krankheitsmodellparameter können im Vorfeld im Rahmen einer Segregationsanalyse geschätzt werden, bei der mittels statistischer Verfahren verschiedene genetische Modelle und Vererbungsmodi mit dem Vererbungsmuster des Krankheitsphänotyps in Familien verglichen werden. Dabei können sowohl Ein-Genort-Modelle, wie sie bei Mendelschen Erkrankungen auftreten, als auch komplexere Modelle, die weitere genetische Faktoren und Umwelteinflüsse berücksichtigen, in die Analyse einbezogen werden (Weiss, 1993). Alternativ können die Krankheitsmodellparameter auch in einer gemeinsamen Kopplungs- und Segregationsanalyse geschätzt werden. Ein Beispiel für ein solches Verfahren ist die MOD-Score-Analyse (Risch, 1984). Durch die Schätzung der Krankheitsmodellparameter im Rahmen der MOD-Score-Analyse kann ein Verlust an statistischer Trennschärfe für den Kopplungstest, wie er bei gewöhnlichen LOD-Score-Analysen durch fehlerhafte Spezifikation der Krankheitsmodellparameter vorkommen kann, reduziert werden (Clerget-Darpoux et al., 1986). Da die Verteilung der Teststatistik der MOD-Score-Analyse im Allgemeinen nicht bekannt ist, müssen Simulationen unter der Nullhypothese keiner Kopplung gemacht werden, um p-Werte für den Kopplungstest zu erhalten. Die Teststatistik der MOD-Score-Analyse ist der Zehnerlogarithmus des Likelihood-Quotienten der Likelihood unter der Alternativhypothese Kopplung und der entsprechenden Likelihood unter der Nullhypothese keiner Kopplung. Die Likelihood L des genetischen Modells, gegeben die beobachteten Daten, ist dabei proportional zur Wahrscheinlichkeit, die Daten zu beobachten, gegeben das Modell. Anders formuliert bezeichnet die Likelihood die Wahrscheinlichkeit, die Daten zu beobachten, unter der Annahme, dass das zugrundeliegende Modell wahr bzw. korrekt spezifiziert ist (Weiss, 1993). Der Likelihood-Quotient wird im Hinblick auf

die Rekombinationsfrequenz (bzw. die genetische Position des Krankheitslocus) und die Krankheitsmodellparameter maximiert:

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

Es ist zu beachten, dass im Zähler und Nenner dieselben Werte für die Krankheitsmodellparameter angenommen werden. Dadurch ist der MOD-Score proportional zur Likelihood, bedingt auf die Krankheitsphänotypen, und liefert unverzerrte Schätzer für die Krankheitsmodellparameter, welche nicht von den Details der Datenerhebung (engl. *ascertainment*) abhängen (Elston, 1989). Dies gilt unter anderem jedoch nur dann, wenn sich die genetischen Marker nicht im LD mit den Allelen am Krankheitslocus befinden (Ginsburg et al., 2004, Malkin und Elston, 2005). Alle Bedingungen, die für die unverzerrte Schätzbarkeit der Krankheitsmodellparameter im Rahmen der MOD-Score-Analyse notwendig sind, wurden in Brugger et al. (2016, Anhang A dieser Arbeit) zusammengetragen. Brugger et al. (2016, Anhang A dieser Arbeit) konnten darüber hinaus anhand von simulierten Familiendaten zeigen, dass mittels der MOD-Score-Analyse eine Schätzung der Krankheitsmodellparameter auch praktisch möglich ist, wobei die Genauigkeit der Schätzung stark sowohl vom zugrundeliegenden Krankheitsmodell als auch von der Komplexität der untersuchten Familien abhängt. Die Berechnung der genetischen Likelihood ist also entscheidend für die Kopplungsanalyse und deren aufwändigster Schritt, vor allem dann, wenn große Familien, viele Marker und fehlende Markergenotypen im Datensatz auftreten. Es sind im Wesentlichen zwei Algorithmen zur Berechnung der Likelihood gebräuchlich: der Elston-Stewart-Algorithmus (Elston und Stewart, 1971) und der Lander-Green-Algorithmus (Lander und Green, 1987). Die Rechenzeit des Elston-Stewart-Algorithmus wächst linear mit der Zahl der Familienmitglieder in einem Stammbaum und exponentiell mit der Zahl der untersuchten Marker. Die Rechenzeit des Lander-Green-Algorithmus hingegen wächst linear mit der Zahl der Marker und exponentiell mit der Zahl der Familienmitglieder in einem Stammbaum, weshalb er gut für die Analyse von Datensätzen mit vielen Markern und kleineren bis mittelgroßen Familien geeignet ist. Der Lander-Green-Algorithmus basiert auf dem Konzept der Vererbungsvektoren, bei dem jeder einzelne Vektor ein mögliches Muster für die Segregation der Allele der Founder in einer Familie an einem Locus beschreibt. Mit n Nonfoundern und f Foundern in einem Stammbaum sind 2^{2n} Vektoren möglich, die sich wiederum in 2^{2n-f} äquivalente Vererbungsvektorklassen zu je 2^f Vektoren zusammenfassen lassen (Kruglyak et al., 1996). Ohne Hinzunahme von Markerinformation entspricht die Wahrscheinlichkeitsverteilung der Vererbungsvektoren (Vererbungsverteilung) an einer gegebenen genetischen Position einer Gleichverteilung. Im Falle von hinreichend vielen informativen Markern ist die Vererbungsverteilung auf wenige Vektoren konzentriert. Mit Hilfe des

Lander-Green-Algorithmus ist nach Ermittlung der Vererbungsverteilung unter Hinzunahme sämtlicher Marker sowohl eine parametrische als auch nicht-parametrische Kopplungsanalyse möglich. Hierzu ist eine entsprechende Scoring-Funktion zur Bewertung der Krankheitsphänotypen zu verwenden.

Das Softwareprogramm GENEHUNTER-MODSCORE (GHM) (Strauch, 2003, Dietter et al., 2007, Mattheisen et al., 2008, Brugger und Strauch, 2014, Brugger et al., 2024) basiert wie das ursprüngliche Softwareprogramm GENEHUNTER (Kruglyak et al., 1996) auf dem Lander-Green-Algorithmus, der zur Berechnung der Vererbungsverteilung eingesetzt wird. Die für eine MOD-Score-Analyse eingesetzte, parametrische Scoring-Funktion beinhaltet die Likelihood für den Krankheitsgenort, gegeben einen bestimmten Vererbungsvektor, welche die Krankheitsmodellparameter, d. h. die Penetranzen und die Krankheitsallelfrequenz, enthält. Diese Krankheitslocus-Likelihood wird für jeden Vererbungsvektor berechnet und mit der aus der Vererbungsverteilung bestimmten Wahrscheinlichkeit für den jeweiligen Vererbungsvektor unter Berücksichtigung sämtlicher Marker gewichtet. Die Beiträge aller Vererbungsvektoren werden aufsummiert und zusätzlich mit der Summe der Krankheitslocus-Likelihoods, die sich bei einer Gewichtung mit einer Gleichverteilung der Vererbungsvektoren (d. h. ohne Markerinformation) ergibt, normiert. Der so berechnete Gesamt-Score ist letztlich ein Likelihood-Quotient, wobei der Zähler der Alternativhypothese Kopplung zwischen Marker- und Krankheitsgenort und der Nenner der Nullhypothese keiner Kopplung entspricht. Der dekadische Logarithmus dieses Likelihood-Quotienten wird über alle Stammbäume im Datensatz summiert, die entsprechend über die Krankheitsmodellparameter maximierte Summe aus logarithmierten Likelihood-Quotienten ergibt dann den MOD-Score. Für eine genetische Position ist die Berechnung der Vererbungsverteilung ein einmaliger Vorgang, die Krankheitslocus-Likelihood muss jedoch für jedes neu eingesetzte Set an Krankheitsmodellparametern im Rahmen des Optimierungsverfahrens, wie es in GHM implementiert ist, neu berechnet werden. Die Berechnung der Krankheitslocus-Likelihood ist somit der geschwindigkeitsbestimmende Schritt einer MOD-Score-Analyse. Brugger und Strauch (2014, Paper I) konnten zeigen, dass es im Rahmen des Lander-Green-Algorithmus möglich ist, Vererbungsvektoren zusammenzufassen, die identische Krankheitslocus-Likelihoods aufweisen, um so die Zahl der nötigen Berechnungen pro eingesetztem Set an Krankheitsmodellparametern zu reduzieren. Dies ist mit Hilfe algebraischer (symbolischer) Berechnungen der Krankheitslocus-Likelihood für jeden Vererbungsvektor und anschließender Zuordnung zu einer Vererbungsvektorklasse mit potenziell mehreren Vererbungsvektoren möglich. Erst nachdem die algebraische Struktur aller Vererbungsvektoren berechnet und Vektoren mit identischer, algebraischer Krankheitslocus-Likelihood zu Vererbungsvektorklassen zusammengeführt worden sind, wird die Krankheitslocus-Like-

lihood durch Einsetzen entsprechender Werte für die Penetranzen und die Krankheitsallelfrequenz ausgerechnet. Dieser Algorithmus kann insbesondere dann zur Verringerung von Rechenzeit einer MOD-Score-Analyse führen, wenn sich mehrere Familien identischer Stammbaumstruktur im Datensatz wiederfinden, da der Algorithmus über Familien hinweg Vererbungsvektorklassen zusammenfassen kann. Ein solches Szenario ist insbesondere für Simulationsstudien gegeben, Anwendungen des in Paper I dargestellten und in GHM implementierten algebraischen Algorithmus sind in Paper III (Anhang A) zur Schätzbarkeit von Krankheitsmodellparametern und IV (Anhang B) zur Untersuchung des genomischen Imprintings zu finden.

2.4 Paper II: Gemeinsame Kopplungs- und Assoziationsanalyse in Familien und Unverwandten

Eine gemeinsame Kopplungs- und Assoziationsanalyse kombiniert Kopplungs- und Assoziationsinformation aus Familien, wobei Assoziationsinformation aus der Allgemeinbevölkerung durch Einbeziehung unverwandter Individuen hinzugefügt werden kann. Dadurch kann die statistische Trennschärfe zur Kartierung eines krankheitsursächlichen Gens speziell dann erhöht werden, wenn der Datensatz aus Familien und Unverwandten besteht (Göring und Terwilliger, 2000). Dies hat zwei Gründe. Zum einen machen Kopplungsanalysen im Allgemeinen die Annahme, dass sich die Allele an Krankheits- und Markerlocus im LE befinden. Es ist jedoch bekannt, dass sich Krankheits- und Markerlocus bei geringer Entfernung auf demselben Chromosom im LD befinden können (Jorde, 1995), wodurch es bei Kopplungsanalysen zu einem Verlust an Trennschärfe kommen kann, wenn das LD nicht berücksichtigt wird (Clerget-Darpoux, 1982). Zum anderen nutzen Assoziationsanalysen diese LD-Information zwar aus, um krankheitsursächliche Gene zu kartieren, sie fällt jedoch schnell mit einer zunehmenden genetischen Distanz zwischen Krankheits- und Markerlocus ab (Terwilliger, 1995). Daher ist die Idee, die im Vergleich zur Assoziationsinformation über größere genetische Distanzen hinweg stabilere Kopplungsinformation in Familien mit der Assoziationsinformation in Familien und Unverwandten in einem gemeinsamen Test zu vereinen, sehr vielversprechend. Erste Ansätze für gemeinsame Kopplungs- und Assoziationsanalysen finden sich in den Arbeiten von MacLean et al. (1984), Clerget-Darpoux et al. (1988) und Tienari et al. (1992). Eine gemeinsame Kopplungs- und Assoziationsanalyse basierend auf dem MOD-Score-Ansatz (siehe Abschnitt 2.3) wird verwirklicht, indem der entsprechende Likelihood-Quotient mit einem Parameter für das LD zwischen Krankheits- und Markerlocus erweitert wird:

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

Die Umsetzung einer solchen Analyse und deren Implementation in GHM ist Gegenstand von Paper II. Die Modellierung des LDs geschieht dabei anhand von bis zu drei SNVs (sogenannten Test-SNVs), deren Allele mit jenen des Krankheitslocus gemeinsame Haplotypen bilden. Dabei tragen lediglich die Founder Beiträge zu den Haplotypfrequenzen bei, Nonfounder dienen jedoch dem Ausschluss von inkompatiblen Haplotypkonfigurationen im Zuge der Haplotypfrequenzschätzung (siehe auch Rohde und Fürst, 2001). Letztere wird in GHM mittels eines Expectation-Maximization-Algorithmus (EM-Algorithmus, Dempster et al., 1977) bewerkstelligt, der im Kontext des Lander-Green-Algorithmus zur Berechnung der auf LD-Parameter erweiterten Krankheitslocus-Likelihood arbeitet (siehe Abecasis und Wigginton, 2005). Kopplungsinformation wird durch flankierende Marker mit beliebiger Zahl von Allelen beigetragen. Die Komplexität des Maximierungsproblems zur Berechnung des MOD-Scores steigt durch die zusätzliche Zahl an Krankheitsmodellparametern gewaltig. So müssen zum Beispiel im Falle von drei SNVs 16 Haplotypfrequenzen im Zähler von Gleichung (2) berücksichtigt werden. Eine solche komplexe Aufgabe ist nicht mehr mit dem bisherigen, numerischen Optimierungsverfahren von GHM in einer überschaubaren Rechenzeit lösbar. Deshalb wurde zur Berechnung des MOD-Scores der Optimierungsalgorithmus COBYLA (Powell, 1994, Powell, 1998) in der Version, wie er in der Funktionsbibliothek NLopt (v2.6.2) (Johnson, 2020) enthalten ist, in GHM implementiert. Der Beitrag allelischer Markerinformation der Founder in Form von Haplotypfrequenzen zum MOD-Score führt dazu, dass die bisherige Simulationsroutine von GHM (Mattheisen et al., 2008) zur Berechnung von empirischen p-Werten für den Kopplungstest nicht für den Test auf gemeinsame Kopplung und Assoziation angewendet werden kann. Daher wurde eine neue Simulationsroutine entwickelt und in GHM implementiert, die bei der Erzeugung von Replikaten unter der Nullhypothese keiner Kopplung und keiner Assoziation die Haplotypfrequenzen der Founder für die Simulation der Markerdaten berücksichtigt. In Paper II wurde diese neue Simulationsroutine umfangreich evaluiert und letztlich validiert. Es ist darüber hinaus direkt mit GHM möglich, die Simulationen auf mehreren Computerkernen parallel laufen zu lassen. Da die Berechnungen für einen empirischen p-Wert für den gemeinsamen Test auf Kopplung und Assoziation sehr aufwändig sind, wird der p-Wert nur für ein bestimmtes, aus maximal drei Test-SNVs bestehendes Markerset berechnet, wohingegen ein explorativer Scan eines ganzen Chromosoms auf gemeinsame Kopplung und Assoziation ohne p-Wert-Berechnung vollautomatisch in GHM unter Angabe einiger, weniger Randbedingungen (z. B. Anzahl der Test-SNVs, maximale Distanz zwischen den Test-SNVs) vollzogen werden kann. Die in Paper II durchgeführte Simulationsstudie hat gezeigt, dass der gemeinsame Test auf Kopplung und Assoziation im Rahmen des MOD-Score-Ansatzes mehr statistische Trennschärfe liefern kann als ein vergleichbarer Test aus dem PSEUDOMARKER-Softwarepaket (Göring und Terwilliger, 2000, Hiekkalinna et al.,

2011, Gertz et al., 2014). Dies ist vor allem bei einem komplexeren Assoziationsmuster der Fall, was bedeutet, dass sich das LD über den Krankheitslocus und zwei bis drei Markerloci erstreckt. Um die neue Methode auch in der praktischen Anwendung zu prüfen, wurden Daten der Nationalen Fallsammlung für familiäres Pankreaskarzinom (FaPaCa) analysiert. Die genetische und molekulare Grundlage des familiären Pankreaskrebses ist weitestgehend unklar, wird aber als sehr komplex angenommen (Bartsch et al., 2021). Die aus den FaPaCa-Familien gewonnenen Markerdaten wurden einem initialen genomweiten Kopplungsscan mittels GHM unterzogen, um vielversprechende Kandidatenregionen für den aufwändigeren gemeinsamen Test auf Kopplung und Assoziation zu selektieren. Die Daten stellen eine große Herausforderung an die Rechenleistung für den gemeinsamen Test auf Kopplung und Assoziation dar, weil die meisten Founder der FaPaCa-Familien untypisiert geblieben sind, wodurch die Berechnung der Krankheitslocus-Likelihood sehr viele mögliche Haplotypkonfigurationen berücksichtigen muss. Als Ergebnis der Analyse konnte eine vielversprechende Region zwischen den Genorten für *IL17REL* und *PIM3* auf Chromosom 22q13.33 identifiziert werden. Der lange Arm von Chromosom 22 steht schon länger im Verdacht, mit der Entstehung von Pankreaskrebs in Verbindung zu stehen (Handel-Fernandez et al., 2000), vielleicht können zukünftige Mutationsanalysen der 22q13.33-Region in FaPaCa-Familien ein weiteres Puzzleteil zum besseren Verständnis der Krankheitsätiologie des familiären Pankreaskrebses beitragen.

3. Zusammenfassung

Die in der vorliegenden Arbeit zusammengefassten Publikationen (Paper I-IV) haben allesamt das Softwareprogramm GENEHUNTER-MODSCORE (GHM) (Strauch, 2003, Dietter et al., 2007, Mattheisen et al., 2008, Brugger und Strauch, 2014, Brugger et al., 2024) und den darin implementierten Ansatz der MOD-Score-Analyse als spezielle Variante der Kopplungsanalyse (Risch, 1984) zum Thema. Alle vier Publikationen hängen thematisch miteinander über GHM und dem darin verwendeten kopplungsanalytischen Ansatz zusammen, weshalb ich sie gerne als meine persönliche „GENEHUNTER-MODSCORE-Tetralogie“ bezeichnen möchte.

Die erste Arbeit der Tetralogie (Paper I, Brugger und Strauch, 2014) beschäftigt sich mit der Frage, wie es gelingen kann, den geschwindigkeitsbestimmenden Schritt einer MOD-Score-Analyse in GHM so zu verändern, dass die Rechenzeit reduziert werden kann. Der im Sinne der Rechenzeit teuerste Schritt in der GHM-MOD-Score-Analyse ist die Berechnung der Krankheitslocus-Likelihood unter Hinzunahme aller Phänotypen der Stammbäume. Diese Likelihood muss für jedes neue zu prüfende Set an Krankheitsmodellparametern, welches im Falle einer dichotomen Erkrankung, die von einem diallelischen Krankheitslocus verursacht wird, aus der Krankheitsallelfrequenz und den Penetranzen besteht, neu ausgerechnet werden. Unter Anwendung des in GHM implementierten Lander-Green-Algorithmus (Lander und Green, 1987) zur Berechnung der Vererbungsverteilung an einer genetischen Position auf Grundlage aller verfügbarer Markerinformation ist es gelungen, eine substantielle Reduzierung der Rechenzeit zu erreichen. Ein gegebener Vererbungsvektor beschreibt ein mögliches Muster für die Segregation der Markerallele der ersten Elterngeneration, der sogenannten Founder, an deren Nachkommen eines Familienstammbaumes. Die Berechnung der Krankheitslocus-Likelihood bezieht ihre Beiträge aus Termen der Krankheitsmodellparameter, die über die verschiedenen, möglichen Vererbungsvektoren summiert werden. Wenn also zwei Vererbungsvektoren denselben Beitrag zur Krankheitslocus-Likelihood liefern, lassen sie sich zu einer Vererbungsvektorklasse zusammenfassen und die Beiträge müssen nicht mehr pro Vektor, sondern nur noch einmal pro Klasse ausgerechnet werden. So kann eine große Menge an Vererbungsvektoren potenziell in einige, wenige Klassen eingeteilt werden, die auch über die Familien hinweg zur Rechenzeiteinsparung genutzt werden können. Dies ist vor allem für Simulationsstudien interessant, bei denen zwar oft eine große Zahl an Familien erzeugt werden, die jedoch auf nur ein paar wenige unterschiedliche Stammbaumtypen zurückgehen. Identische Stammbaumstrukturen liefern hernach dieselben Vererbungsvektorklassen, wodurch eine weitere Optimierung der Rechenzeit erfolgen kann.

Diese algorithmische Optimierung der Rechenzeit für GHM ermöglichte es dann, die in Brugger et al. (2016, Paper III, Anhang A) beschriebene Simulationsstudie zur Untersuchung der Schätzbarkeit von Krankheitsmodellparametern im Rahmen einer MOD-Score-Analyse durchzuführen. Hierzu wurde als Vorarbeit ein sorgfältiger Review darüber geführt, inwieweit und unter welchen Bedingungen die Krankheitsmodellparameter im Rahmen einer MOD-Score-Analyse theoretisch überhaupt schätzbar sind, wobei vor allem auf die Arbeiten von Elston (1989), Ginsburg et al. (2004) sowie Malkin und Elston (2005) verwiesen werden muss. Als Ergebnis der Simulationsstudie konnte festgestellt werden, dass Krankheitsmodellparameter im Rahmen einer MOD-Score-Analyse in der Tat praktisch schätzbar sind, deren Schätzgenauigkeit und Identifizierbarkeit aber stark von den wahren Krankheitsmodellparametern und den untersuchten Stammbaumstrukturen abhängen.

Eine weitere Simulationsstudie, die in Brugger et al. (2019, Paper IV, Anhang B) beschrieben ist, hat das Confounding zwischen geschlechtsspezifischen Rekombinationsfrequenzen und genomischem Imprinting im Rahmen von Kopplungsanalysen mit GHM zum Thema. Hierzu wurde auch ein neuer Test auf Imprinting auf Basis des MOD-Scores („MOBIT“) vorgeschlagen und dessen statistische Eigenschaften untersucht. Wenn die Expression eines krankheitsursächlichen Gens von dessen elterlicher Herkunft abhängt, so spricht man von genomischem Imprinting (Hall, 1990), welches durch die Aufspaltung der Heterozygotenpenetranz in zwei nach der elterlichen Herkunft des Krankheitsallels getrennten Penetranzen im Rahmen einer LOD- bzw. MOD-Score-Analyse berücksichtigt werden kann (Strauch et al., 2000). Der MOBIT als Test auf Imprinting ergibt sich dann als Differenz des MOD-Scores unter Berücksichtigung des Imprintings und des MOD-Scores ohne Berücksichtigung des Imprintings. Alternativ lässt sich genomisches Imprinting im Rahmen einer Kopplungsanalyse auch über geschlechtsspezifische Rekombinationsfrequenzen modellieren (Smalley, 1993), was zu einer Confoundingsituation führen kann, wenn auf Imprinting in Gegenwart von tatsächlich vorliegenden geschlechtsspezifischen Rekombinationsfrequenzen getestet wird, diese aber nicht in der Analyse berücksichtigt werden. Durch umfangreiche Simulationen konnte in Paper IV gezeigt werden, dass sich Confounding vermeiden lässt, wenn man einen Multimarkeransatz wählt, dessen Markerabstände unter 1 cM liegen. In Paper IV wurden die asymptotische Verteilung unter der Nullhypothese keinen Imprintings, aber Kopplung angegeben und zwei Simulationsroutinen vorgeschlagen, mit denen man empirische p-Werte für den MOBIT erhalten kann.

Das letzte Paper der GHM-Tetralogie (Brugger et al., 2024, Paper II) beschäftigt sich abschließend mit einer Erweiterung des MOD-Score-Ansatzes, der es ermöglicht, einen gemeinsamen Test auf Kopplung und Assoziation in Familien und Unverwandten

durchzuführen. Der MOD-Score wurde hierfür um einen Parameter für das LD erweitert, welcher in Form von Haplotypfrequenzen der Founder und der Unverwandten unter Hinzunahme von bis zu drei SNVs zusätzlich zum Krankheitsgenort parametrisiert wurde. Die stark vergrößerte Zahl zu optimierender Parameter im entsprechenden Likelihood-Quotienten des MOD-Scores machte die Implementation eines Optimierungsalgorithmus in GHM notwendig (COBYLA (Powell, 1994, Powell, 1998)), ohne den die Rechenzeit für MOD-Score-Analysen nicht mehr handhabbar gewesen wäre. Um einen empirischen p-Wert für den MOD-Score unter der Nullhypothese keiner Kopplung und keiner Assoziation zu erhalten, wurde eine neue Simulationsroutine implementiert und erfolgreich anhand ausführlicher Simulationen validiert. Die statistischen Eigenschaften (Fehler 1. Art, statistische Trennschärfe) des neuen MOD-Scores wurden ebenfalls ausführlich mittels Simulationen untersucht und mit einem vergleichbaren Test aus dem Softwareprogramm PSEUDOMARKER (Göring und Terwilliger, 2000, Hiekkalinna et al., 2011, Gertz et al., 2014) verglichen. Dabei konnte festgestellt werden, dass der gemeinsame Test auf Kopplung und Assoziation auf Basis des MOD-Scores vor allem dann eine höhere Trennschärfe aufweist, wenn sich das LD über den Krankheitslocus und zwei bis drei SNVs erstreckt. Die Analyse der Daten der FaPaCa-Familien erbrachte einen vielversprechenden, signifikant gekoppelten und assoziierten Locus auf Chromosom 22q13.33, welcher in zukünftigen Mutationsanalysen zum besseren Verständnis der Krankheitsätiologie des familiären Pankreaskrebses beitragen kann.

4. Abstract

All publications comprised in this work deal with the GENEHUNTER-MODSCORE (GHM) software package (Strauch, 2003, Dietter et al., 2007, Mattheisen et al., 2008, Brugger und Strauch, 2014, Brugger et al., 2024) and the MOD score approach in linkage analysis (Risch, 1984) implemented therein. The publications hence represent what I would like to call my personal ”GENEHUNTER-MODSCORE tetralogy”.

The first publication (Brugger and Strauch, 2014, Paper I) concerns the question as how to speed up the most time-consuming step in a MOD score analysis, which is the calculation of the disease-locus likelihood. The disease-locus likelihood makes use of the disease phenotypes of all family members in a pedigree and has to be recalculated for every new tested set of trait-model parameters, i.e., the disease allele frequency and the penetrances for a dichotomous trait governed by a diallelic disease locus. In the context of the Lander-Green algorithm (Lander and Green, 1987), which is employed to calculate the inheritance distribution at a given genetic position using all available marker information, we were able to speed up the calculations substantially. Specifically, an inheritance vector denotes a possible segregation pattern of founder alleles to their offspring in a given pedigree. The disease-locus likelihood calculation entails the summation over contributions of all possible inheritance vectors, whereby each inheritance vector contributes terms according to the specified trait-model parameters. If two distinct inheritance vectors yield the same contribution to the disease-locus likelihood, they can be grouped together into an inheritance vector class. This way, potentially many inheritance vectors can be grouped into fewer inheritance vector classes. Hence, the contributions to the disease-locus likelihood only need to be calculated once for each class instead of each vector. It can furthermore be shown that inheritance vector classes can even be used across pedigrees to save computation time, which is especially relevant for linkage simulation studies. Typically, a large number of families are generated in simulation studies, however, they usually correspond to a few distinct pedigree types. Because identical pedigrees (including each member’s phenotype, but not genotype) lead to the same set of inheritance vector classes, calculation time can be saved even across pedigrees.

The implementation of the above-described algorithmic optimization in GHM facilitated the investigation of the capacity of a MOD score analysis to estimate trait-model parameters by performing a simulation study (Brugger et al., 2016, Paper III, Anhang A). To this end, theoretical arguments concerning the ability of the MOD score approach to estimate trait-model parameters were carefully reviewed (see e.g. Elston (1989), Ginsburg et al. (2004) as well as Malkin and Elston (2005)). As a result, the simulation study showed that trait-model parameters can in fact be estimated in practice using the MOD score approach, however, the accuracy of the estimates and the identifiability

of the parameters strongly depend on the truly underlying trait-model parameters and the pedigree types used in the analysis.

In a further simulation study (Brugger et al., 2019, Paper IV, Anhang B), the confounding between sex-specific recombination fractions and genomic imprinting was systematically investigated. To this end, a new test statistic to test for the presence of imprinting in the context of linkage analysis using MOD scores (“MOBIT”) was proposed and its statistical properties were thoroughly evaluated. Genomic imprinting means the dependence of an individual's liability to develop a disease according to the parental origin of the mutated allele(s) (Hall, 1990). In the context of parametric linkage analysis, imprinting can be modelled by splitting up the heterozygote penetrances into two penetrances according to the parental origin of the mutated allele (Strauch et al., 2000). The MOBIT is defined as the difference between the MOD score accounting for imprinting and the MOD score not accounting for imprinting. Alternatively, in the context of linkage analysis, imprinting can be modelled using sex-specific recombination fractions (Smalley, 1993). Hence, if sex-specific recombination fractions are truly present, but not accounted for in the analysis, this can lead to false-positive results of linkage-based imprinting tests, i.e., confounding. The results of the extensive simulation study in Paper IV showed that confounding for the MOBIT can be avoided using a multi-marker approach, with markers spaced less than 1 cM from each other. Furthermore, the asymptotic distribution of the MOBIT under the null hypothesis of linkage, but no imprinting was presented, together with two proposed simulation strategies to obtain empiric p values for the MOBIT.

The final paper of the GHM tetralogy (Brugger et al., 2024, Paper II) describes an extension to the MOD score approach that enables a joint linkage and association analysis using families and unrelated individuals. To this end, the MOD score was extended to include a parameter for LD, which was parametrized in terms of founder haplotype frequencies between alleles at the disease locus and up to three SNVs. Due to the increased number of parameters that have to be optimized in the likelihood ratio of the extended MOD score, the derivative-free optimization algorithm COBYLA (Powell, 1994, Powell, 1998) was implemented in GHM to speed up calculation time. In addition, a novel simulation routine to obtain empiric p values for the joint linkage and association test was implemented in GHM and validated using simulated data. The statistical properties of the extended MOD score, i.e., type I error and power, were also evaluated using an extensive simulation study and compared to a commonly used joint linkage and association test implemented in the PSEUDOMARKER software package (Göring and Terwilliger, 2000, Hiekkalinna et al., 2011, Gertz et al., 2014). As a result, it could be shown that the joint linkage and association test using the MOD score approach outperforms the PSEUDOMARKER test in terms of power for scenarios, in

which LD ranges over the trait locus and more than one, i.e., two to three, SNVs. To evaluate the extended joint linkage and association MOD score in practice, pedigree data from the FaPaCa registry were analyzed as a use case. Consequently, the analysis revealed a promising locus on chromosome 22q.13.33, which could serve as a candidate for mutation analysis to further elucidate the disease etiology of familial pancreatic cancer.

5. Paper I

Brugger M, Strauch K. Fast linkage analysis with MOD scores using algebraic calculation. Hum Hered. 2014;78(3-4):179—194.

Original Paper

Human Heredity

Hum Hered 2014;78:179–194
DOI: 10.1159/000369065

Received: June 23, 2014
Accepted after revision: October 14, 2014
Published online: January 29, 2015

Fast Linkage Analysis with MOD Scores Using Algebraic Calculation

Markus Brugger Konstantin Strauch

Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, and Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

Key Words
Parametric linkage analysis · MOD scores · Lander-Green algorithm · Algebraic calculation · Rare variants

Abstract
Objective: As the mode of inheritance is often unknown for complex diseases, a MOD-score analysis, in which the parametric LOD score is maximized with respect to the trait-model parameters, can be a powerful approach in genetic linkage analysis. Because the calculation of the disease-locus likelihood is the most time-consuming step in a MOD-score analysis, we aimed to optimize this part of the calculation to speed up linkage analysis using the GENEHUNTER-MODSCORE software package. ***Methods:*** Our new algorithm is based on minimizing the effective number of inheritance vectors by collapsing them into classes. To this end, the disease-locus-likelihood contribution of each inheritance vector is represented and stored in its algebraic form as a symbolic sum of products of penetrances and disease-allele frequencies. Simulations were used to assess the speedup of our new algorithm. ***Results:*** We were able to achieve speedups ranging from 1.94 to 11.52 compared to the original GENEHUNTER-MODSCORE version, with higher speedups for larger pedigrees. When calculating p values, the speedup ranged from 1.69 to 10.36. ***Conclusion:*** Computation times for MOD-score analysis, involving the evaluation of many tested sets of trait-model parameters and p value calculation, have been prohibitively high so far. With our new algebraic algorithm, such an analysis is now feasible within a reasonable amount of time.

Introduction

Since its first successful application by the physician and geneticist Jan Mohr in 1954 [1], linkage analysis has been a powerful tool in human disease gene mapping for many decades. With this method, many Mendelian disease genes have been mapped to their genetic loci by the use of family data [2]. Due to the development of genotyping techniques with dense SNP marker panels and the progressing availability of large case-control or population-based cohorts, association analysis has recently become the preferred method for statistical analysis in the field of genetic epidemiology. Unlike linkage analysis, an association analysis can make use of samples with unrelated individuals; it does not require families which are obviously much harder to recruit. However, with the advent of next-generation sequencing data and increasing interest in the analysis of rare variants, the analysis of family data using linkage analysis is undergoing a renaissance. The basis for this interest is that numerous rare

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

0001–5652/15/0784–0179$39.50/0

Markus Brugger
Institute of Genetic Epidemiology, Helmholtz Zentrum München
German Research Center for Environmental Health
Ingolstädter Landstrasse 1, DE–85764 Neuherberg (Germany)
E-Mail markus.brugger@helmholtz-muenchen.de

variants with moderate effects may explain an appreciable amount of the missing heritability [3]. Although rare variants are individually rare, a single person can have thousands of such rare variants across the genome. It can thus be difficult to determine whether the observation of a rare variant is a sequencing artifact or in fact a true variant if it is carried by only a single individual of the sample. However, one expects that rare variants segregate and accumulate within families. Results from the Genetic Analysis Workshop 17 showed that analyses using whole-exome sequencing data require much smaller sample sizes when working with families than with unrelated individuals, because the ability to detect rare causal variants is enhanced in family studies as the variants are carried by several family members jointly [4].

In parametric linkage analysis, which is also known as LOD-score or model-based analysis, a certain set of trait-model parameters is explicitly assumed for the segregation of the disease. In the simplest case of a diallelic autosomal trait locus, which is assumed throughout this paper, these parameters are the disease-allele frequency p 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. The central part of parametric linkage analysis is the computation of the genetic likelihood, which is based on the following parameters: disease-allele frequency, penetrances, marker-allele frequencies, and the recombination fractions – and, if applicable, linkage disequilibria between the loci. In addition, the relation between family members is required to be known. Eventually, a likelihood-ratio test is performed, in which the likelihood under the alternative hypothesis of linkage with some specific value of the recombination fraction ($\theta < 0.5$; the numerator of the likelihood ratio) is compared to the null hypothesis of no linkage ($\theta = 0.5$; the denominator of the likelihood ratio). The logarithm to the base 10 of this likelihood ratio is the LOD score [5]. It is maximized by varying θ between marker and trait locus in the numerator (maximum LOD score). 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. The latter approach is also known as MOD-score analysis and has been first proposed by Risch [6]. As the power of a LOD-score analysis crucially depends on the true mode of inheritance, which is generally unknown, a MOD-score analysis can have greater power to detect linkage than a simple LOD-score analysis. Furthermore, in case of a trait-model-parameter misspecification, the recombination fraction will be overestimated [7]. In a multipoint analysis, the misspecification may even lead to an exclusion of linkage [8]. Simulations have shown that, especially when analyzing a mixture of different types of pedigrees, the MOD-score approach outperforms other linkage methods in terms of power to identify genes with modest effect [9]. 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. Our group has implemented the MOD-score approach, including a routine to perform simulations under the null hypothesis, in the GENEHUNTER-MODSCORE (GHM) software [10–13]. Its application has led to the identification of a variety of genetic disease loci [14–18].

Nonparametric linkage methods have been proposed in order to avoid trait-model misspecification that occurs when using simple LOD-score analyses. These methods test if affected pedigree members have more alleles in common than would be expected by chance under the null hypothesis of no linkage. Nonparametric methods are often considered to be 'model-free' because they do not rely on explicit assumptions as to the trait-model parameters. However, Knapp et al. [19] have shown that, for samples of affected sib pairs (ASPs) with the parents' phenotypes unknown or set to unknown, the nonparametric mean test is equivalent to a LOD-score analysis under a recessive mode of inheritance, and the possible triangle test proposed by Holmans [20] is equivalent to a MOD-score analysis. In the possible triangle test, the genetic likelihood is expressed in terms of the probabilities z_0, z_1, and z_2 that an ASP shares 0, 1, or 2 alleles identical-by-descent (IBD) with restrictions to genetically possible models [20]. These allele-sharing probabilities can be expressed as functions of the trait-model parameters f_0, f_1, f_2, p, and θ [21], and hence, the parametric and nonparametric likelihood are identical. More generally, the allele-sharing probabilities of any pedigree with affected relatives could be used to construct a nonparametric allele-sharing-based test statistic [22]. However, for such a nonparametric test to be constructed for a certain pedigree type other than ASPs or affected half-sib pairs (AHSPs) would yet demand knowledge as to how many allele-sharing classes exist for that pedigree type and how the corresponding restrictions to genetically possible models can be formulated. Knapp [23] derived allele-sharing probabilities for affected sib triplets (ASTs) with parental phenotypes set to unknown. However, the re-

strictions to genetically possible models cannot be expressed in closed form. But again, the allele-sharing probabilities, which represent the truly underlying parameters, can be modeled as a function of f_0, f_1, f_2, p, and θ. Hence, the parametric and nonparametric likelihood are identical even beyond the special cases of ASPs and AHSPs, and MOD-score analysis is equivalent to the likelihood-ratio test based on allele-sharing parameters. As outlined by Strauch [22], this holds for any type of pedigree.

The calculation of the genetic likelihood is pivotal for both parametric and nonparametric linkage analysis. Given the complexity of real family data, it cannot be calculated manually in most cases. Large pedigrees, many markers, and missing genotypes lead to a substantial number of possible genotype combinations that must be considered in the likelihood. Two major algorithms are known that allow for the calculation of the likelihood: the Elston-Stewart [24] and the Lander-Green algorithm [25]. The former is genotype-oriented and is based on the peeling of nuclear families. It makes use of the independence of genotypes of different nuclear families within a pedigree when conditioning on a certain genotype of the connecting person, the so-called pivot. The Elston-Stewart algorithm thereby summarizes identical terms that correspond to a particular genotype combination within the likelihood. The algorithm scales linearly with the number of individuals in a pedigree and exponentially with the number of analyzed loci. Hence, it is limited to the analysis of a relatively small number of genetic markers. The Elston-Stewart algorithm has been implemented and further optimized in several linkage software packages such as LINKAGE [26–28], FASTLINK [29, 30], VITESSE [31, 32], and PSEUDOMARKER [33, 34]. The Lander-Green algorithm is complementary to the Elston-Stewart algorithm, such that it treats each marker locus one after another and distinguishes the marker loci from the disease locus. The Lander-Green algorithm is implemented in several genetic analysis software packages such as GENEHUNTER [35], ALLEGRO [36, 37], and MERLIN [38]. It scales linearly with the number of markers and exponentially with the number of individuals in a pedigree. Therefore, the Lander-Green algorithm is well suited for the analysis of large datasets of genetic markers, which are typically available for small to moderately large pedigrees when mapping complex-disease genes. In addition, it allows both parametric and nonparametric linkage analysis. This is because, as a first step, inheritance information is extracted solely from marker data by applying the concept of inheritance vectors. Then, a parametric or nonparametric scoring function that incorporates information with regard to the disease phenotypes of the pedigree members is applied to evaluate a set of genetic positions of the putative trait locus in terms of linkage with the markers. In the parametric case, the scoring function corresponds to the ratio of the disease-locus likelihoods under the assumption of linkage versus no linkage.

In this paper, we describe a new algorithm for the calculation of the parametric disease-locus likelihood in the context of the Lander-Green algorithm. This part of the calculation is the most time-consuming step in a MOD-score analysis. How can it be accelerated? Our new approach to a faster implementation is structured according to the following three aspects:

- *Inheritance Vectors and the Identity of the MOD Score with the Allele-Sharing-Based Test Statistic*. Inspired by the identity of the allele-sharing-based nonparametric likelihood and the parametric likelihood in the test for linkage, our new algorithm is based on minimizing the effective number of inheritance vectors by collapsing them into classes, whose members are observed with the same probability function of f_0, f_1, f_2, and p, i.e. having the same allele-sharing proportions for a given type of pedigree structure. This approach has the potential to considerably reduce the number of floating number operations, because instead of calculating the disease-locus-likelihood contribution for a given set of trait-model parameters for each inheritance vector, it needs to be calculated only once for all members of a certain class.
- *Algebraic Formulation of the Disease-Locus Likelihood*. To collapse inheritance vectors into certain classes, i.e. to recognize which vectors belong to the same class, the disease-locus-likelihood contribution of each inheritance vector must be represented and stored in its algebraic form. This involves representing it as a symbolic sum of products of penetrances and disease-allele frequencies for a given combination of disease-locus genotypes of all individuals in the pedigree. Inheritance vectors with identical symbolic sums can thus readily be grouped into the same class. This step involves no numerical calculation and needs to be done only once at the beginning of a MOD-score analysis for a given pedigree.
- *Exploiting Similarities in Family Structures by the Use of Inheritance Vector Classes*. Two pedigrees with a certain pattern of disease status, each of which can be represented by a directed acyclic graph, are indistin-

guishable in terms of the disease-locus-likelihood structure if they are comprised of the same set of inheritance vector classes and the same number of vector members per class. Hence, two such pedigrees yield the same disease-locus-likelihood contributions. The computational effort for LOD-score calculation for the second pedigree can be entirely avoided. When two pedigrees are distinct, i.e. yielding different sets of inheritance vector classes, identical symbolic products are still stored in a common database to avoid dispensable numerical calculations. The computational effort during the LOD-score calculation is hence further reduced by the degree of similarity of pedigrees based on their inheritance vector classes.

In conjunction with the already existing options and optimizations of GHM, which are addressed below, our new algorithm allows for a rapid evaluation of the likelihood for a large number of disease models, as required during maximization over trait models in a MOD-score analysis. The reduction of computing time is a prerequisite for empirically determining p values by performing simulations and MOD-score calculations of many replicates.

It has to be noted that the first version of GHM [13] is based on GENEHUNTER version 2.1 [39]. Since the release of GENEHUNTER version 1.0 in 1996 [35], many improvements have been implemented, which have led to a significant analysis speedup and which have added various additional functionalities to the software package [39–41]. However, these previous improvements did not concern the calculation of the parametric disease-locus likelihood as does our new algebraic algorithm. All improvements as of GENEHUNTER version 2.1 [39] have been carried forward to GHM and are complementary to the algebraic algorithm presented in this paper. For more information on the original GENEHUNTER software, we refer to the review by Nyholt [42].

Methods

The Lander-Green Algorithm

Inheritance Vectors

As a first step, the Lander-Green algorithm enumerates all possible inheritance vectors in a pedigree. An inheritance vector denotes a possible family-specific pattern of segregation of founder alleles. Each bit of the inheritance vector corresponds to the outcome of a certain meiosis, which codes the transmission of the grand-paternally or grand-maternally inherited allele to the child as a value of 0 or 1, respectively. With n non-founders, there are $2n$ meioses and 2^{2n} possible inheritance vectors. However, even if the information is complete, there are 2^f remaining inheritance vectors that all have the same probability. This is due to the fact that the parental origin of founder haplotypes is unknown. In other words, the bit corresponding to the first child of each founder can be fixed arbitrarily (e.g. to a value of 0). Hence, the 2^{2n} inheritance vectors can be grouped into $2^{(2n-f)}$ equivalence classes, each comprising 2^f inheritance vectors.

Probability of Observed Marker Genotypes Given a Particular Inheritance Vector

The algorithm iterates over inheritance vectors and markers and calculates the probability of the observed genotypes for each marker conditional on a particular inheritance vector [25]. This step of the calculation is based on a graph-theoretical process. Following the notation in Kruglyak et al. [35], let $G(v)$ be a graph for a given inheritance vector v whose vertices are the founder alleles $X = \{x_1, x_2, \ldots, x_{2f}\}$ corresponding to the $2f$ founder alleles at the marker locus, which are assumed to be distinct by descent ('placeholder alleles'). An inheritance vector v specifies the placeholder alleles inherited by each individual in the pedigree. The lines connecting the two placeholder alleles that correspond to the genotype of each individual, as defined by the inheritance vector v, represent the edges of the graph. The placeholder alleles are then assigned the actual founder alleles at the marker locus, and placeholder allele assignments that are incompatible with the observed marker genotypes are eliminated from further consideration. Then, the probability of drawing the founder alleles from the population, i.e. the product of allele frequencies of all founders, is calculated, and the sum of this product is taken over all possible founder allele assignments that are compatible with both the inheritance vector and the observed marker genotypes.

The Markov Chain

The Lander-Green algorithm uses a Markov process to describe the joint distribution of inheritance vectors along a chromosome [25]. This is based on the observation that, under the assumption of no genetic interference, inheritance vectors form a hidden Markov chain. The observed states are the typed marker genotypes, and the hidden states are the inheritance vectors. The matrices of transition probabilities between inheritance vectors at consecutive markers are a function of recombination fractions between markers. After the inheritance vector distribution ($P_{complete}$) has been calculated at a certain genetic position, the disease phenotypes of the family members are considered by using an appropriate scoring function.

The Scoring Function

At this stage of the analysis, different scoring functions are defined for parametric and nonparametric linkage analysis. In a parametric analysis, the scoring function is the ratio of the disease-locus likelihoods under linkage in the nominator versus under no linkage in the denominator. The disease-locus likelihood is calculated conditional on each inheritance vector. As marker information is often incomplete, several inheritance vectors are possible, and the conditional probabilities of these vectors given the marker information ($P_{complete}$) have a nonzero value. Therefore, the sum of the scoring function is taken over all inheritance vectors weighted by their conditional probability given the marker information ($P_{complete}$). Under no linkage between marker and disease locus, the probability of each inheritance vector no longer depends on the marker data. Hence, the inheritance vector distribution at a putative disease lo-

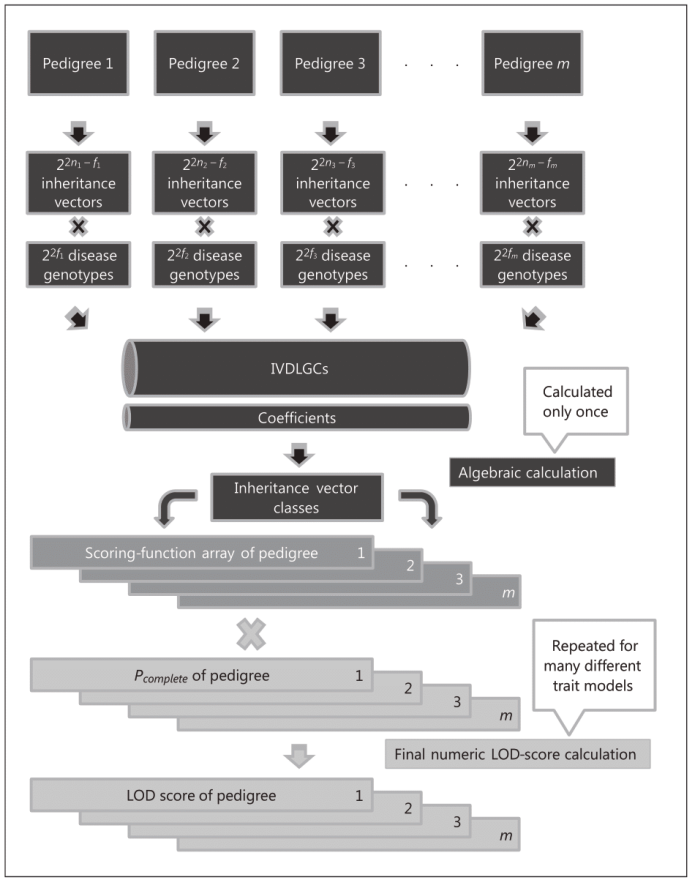


Fig. 1. Depiction of the algebraic algorithm. Steps that have to be calculated only once are highlighted in black. The final LOD-score calculation is shaded in light grey and the interface between the algebraic algorithm and the numeric LOD-score calculation – the scoring-function arrays of the pedigrees – is shown in dark grey. Each inheritance vector of a given pedigree with n nonfounders and f founders is analyzed in regard to its disease-locus-likelihood contribution. For a given inheritance vector, all possible disease-locus-genotype combinations must be considered. Each disease-locus-genotype combination yields a likelihood contribution that is a product of penetrances and disease-allele frequencies. The sum over all disease-locus-genotype combinations is the total disease-locus-likelihood contribution of the given inheritance vector. The likelihood contribution of each IVDLGC is stored in its algebraic form. IVDLGCs of a given inheritance vector that lead to the same algebraic representation are joined together by including a coefficient. Inheritance vectors with the same set of IVDLGCs are assigned to a certain inheritance vector class. The analysis of inheritance vectors is performed for all pedigrees of the dataset, whereby all pedigrees of the sample have a joint IVDLGC storage. This way, a certain inheritance vector class can comprise inheritance vectors of several pedigrees. Finally, the trait-model-specific LOD score is calculated numerically as the scalar product of $P_{complete}$ and the scoring-function array. This step is repeated many times during a MOD-score analysis by numerically evaluating the scoring-function arrays assuming different sets of trait-model parameters.

cus position unlinked to the marker locus corresponds to a uniform distribution with probability $1/2^{(2n-f)}$ for each inheritance vector. Maximizing the logarithm to the base 10 of this likelihood ratio over the recombination fraction θ yields the LOD score. When it is maximized over (f_0, f_1, f_2, and p) in addition to θ, the MOD score is obtained. Nonparametric scoring functions count the number of alleles shared IBD by affected pedigree members given a certain inheritance vector. Popular nonparametric scoring functions are S_{pairs} and S_{all} [35, 43]. Our new algorithm only affects the calculation of the parametric scoring function, and we refer to McPeek [44] for more information about nonparametric scoring functions.

The Algebraic Algorithm

Basic Concept

As described by Strauch [22], inheritance vectors can be collapsed into inheritance vector classes if they cannot be distinguished from each other on the basis of the phenotypic structure of a given family tree. In other words, inheritance vectors being observed with the same allele-sharing probability z_i conditional on the disease phenotypes and the parameters f_0, f_1, f_2, and p are comprised in a certain inheritance vector class. The number of inheritance vector classes, and hence allele-sharing probabilities, depends on the number of persons in a pedigree and hence differs between different types of pedigrees in a sample. As stated before, it appears to be very difficult to construct a nonparametric allele-sharing test, which uses the probabilities z_i, along the lines of the possible triangle test for ASPs, for each of the various pedigree types contained in the particular sample under study. In addition, the restriction to genetically possible models is difficult to formulate. However, given the identity of the parametric likelihood with the nonparametric likelihood in an allele-sharing-based test and the consequential fact that the z_is are a function of (f_0, f_1, f_2, and p),

it seems straightforward to use the parametric formulation of the disease-locus likelihood and to collapse those inheritance vectors into a certain class that, by an identical probability z_i, lead to the same likelihood contribution. An algorithm that makes use of this structure has the potential to substantially reduce the computational effort involved in the disease-locus-likelihood calculation for a given pedigree, since the likelihood needs to be calculated only for one member of each class.

Analysis of Inheritance Vectors

Our new algorithm starts by analyzing each of the $2^{(2n-f)}$ inheritance vectors of a certain pedigree with regard to its disease-locus-likelihood contribution. The processing of the marker-locus likelihood by the GHM software using hidden Markov models to calculate $P_{complete}$ remains untouched by our new approach. The consecutive steps of the algebraic algorithm can be followed by looking at figure 1, which depicts the analysis of all pedigrees in a dataset. For the present, we assume that there is only a single pedigree in the dataset. For a given inheritance vector, all possible disease-locus-genotype combinations must be considered. Each disease-locus-genotype combination yields a likelihood contribution that is a product of penetrances and disease-allele frequencies. The sum over all disease-locus-genotype combinations is the total disease-locus-likelihood contribution of the given inheritance vector. In order to avoid many floating point operations each time an inheritance-vector-disease-locus-genotype combination (IVDLGC) is considered, every IVDLGC is stored in its algebraic form. This way, each inheritance vector can be considered as a set of a certain number of IVDLGCs, whereby our algorithm builds up a database of IVDLGCs, such that only combinations leading to a new algebraic representation are additionally stored in memory. Essentially, IVDLGCs are stored in a big table and connected to the inheritance vector classes by the use of pointers. Pointers are a powerful feature for memory access specific to the C programming language, in which GHM is written. IVDLGCs of a given inheritance vector that lead to the same algebraic representation, i.e. the product of a certain combination of parameters (f_0, f_1, f_2, and p), are joined together by incrementing a coefficient (integer) and thus need not be saved separately, which avoids extra floating point operations and memory.

Identification of Inheritance Vector Classes

All inheritance vectors of a certain class consist of the same set of IVDLGCs. In particular, if an inheritance vector has the same set of IVDLGCs as an inheritance vector class already identified during the course of the calculation, the vector is added to that class. A previously unobserved set of IVDLGCs for a certain vector leads to the definition of a new inheritance vector class. An inheritance vector class corresponds to a certain allele-sharing class in the nonparametric context. Figure 2 gives a technical depiction of the algebraic algorithm for an AST. It illustrates how a specific inheritance vector is assigned to its corresponding class on the basis of the algebraic calculation of its disease-locus-likelihood contribution.

Calculation of the LOD Score

When all inheritance vectors of a given pedigree have been assigned to a certain inheritance vector class and the algebraic structure mentioned above has been determined, the LOD score can readily be calculated for a given set of trait-model parameters. To this end, the algebraic representations of IVDLGCs of all inheritance vector classes are evaluated numerically by inserting the (numeric) values of the parameters (f_0, f_1, f_2, and p) of a specified disease model. The result of each of these products is further multiplied by its associated coefficient, which is equal to the number of IVDLGCs with the same product in a given inheritance vector class, and the sum is taken over all products of that class. This way, the disease-locus-likelihood contributions of all inheritance vector classes are calculated in a single step and then copied into the scoring-function array of the pedigree, according to the class to which a certain inheritance vector belongs. The step of finding the disease-locus-likelihood contribution of the inheritance vector class that corresponds to a given inheritance vector involves the use of pointers and dereference operations. Finally, the trait-model-specific LOD score is calculated as the scalar product of $P_{complete}$ and the scoring-function array. It is of note that information from marker data only affects the calculation of $P_{complete}$, which furthermore is independent of the trait-model parameters. Consequently, $P_{complete}$ has to be computed once for every genetic position and every pedigree in the dataset, even if some or many pedigrees have the same structure. However, $P_{complete}$ can be reused for the LOD-score evaluations under many different trait-model parameters during the maximization.

Number of Inheritance Vector Classes

The degree to which inheritance vectors can be collapsed into certain inheritance vector classes, and hence the computational speedup, depends on the pedigree size and the phenotypes of its members. For example, with nuclear families and parental phenotypes unknown, the potential of reduction by collapsing inheritance vectors into classes increases from ASPs over ASTs to larger sibships. ASPs with 4 possible inheritance vectors have 3 distinct allele-sharing classes, i.e. inheritance vector classes (0, 1, or 2 alleles shared IBD). If imprinting is modeled, e.g. using the four-penetrance formulation developed by Strauch et al. [45] as implemented in GENEHUNTER-IMPRINTING and GHM, ASPs have 4 allele-sharing classes (in this case, the class of 1 shared allele is further distinguished by the parental origin). ASTs with 16 possible inheritance vectors have 4 and 5 allele-sharing classes for a nonimprinting and an imprinting model, respectively (Appendix) [23]. In the following, we will assume an imprinting model when deriving allele-sharing classes, because GHM internally always uses the four-penetrance formulation. The total number of inheritance vectors as well as the reduced number of vector classes are given in table 1 as a function of sibship size of a nuclear family with parental phenotypes unknown (or set to unknown).

Extension across Pedigrees

A further advantage of the algebraic algorithm is that the concept of storing IVDLGCs can even be extended across pedigrees, such that all pedigrees of the sample have a joint IVDLGC storage. A pedigree can thus be considered as a set of certain inheritance vector classes each consisting of a certain set of IVDLGCs. This structure, which is the basis of the algebraic algorithm, is depicted in figure 1. Here, in contrast to the case of considering a single pedigree, a certain inheritance vector class can comprise inheritance vectors of several pedigrees. Hence, the disease-locus-likelihood contributions of all inheritance vector classes are calculated in a single step for the entire dataset, and then the result for a certain inheritance vector class is used for all pedigrees with inheritance vectors that are members of that particular class.

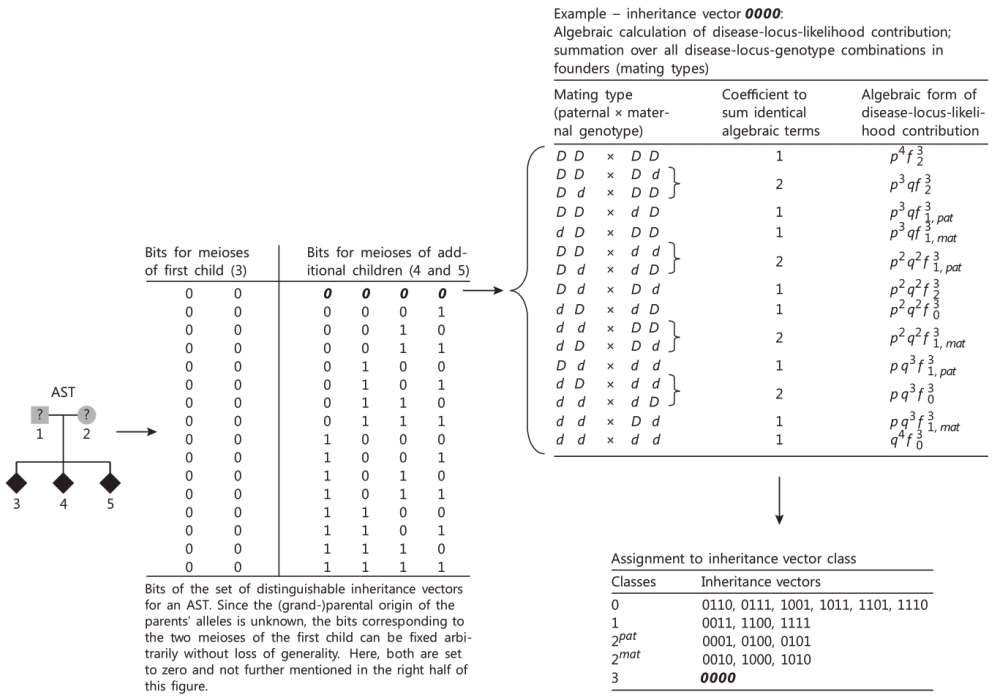


Fig. 2. Technical depiction of the algebraic algorithm for an AST. If several inheritance vectors have the same disease-locus-likelihood contribution, they are joined together in an inheritance vector class.

Table 1. Allele-sharing classes for affected sibships

	ASP	AST	ASQ	ASQui	ASS
Number of inheritance vectors ($2^{(2n-f)}$)	4	16	64	256	1,024
Inheritance vector classes with imprinting taken into account	4	5	11	14	24
Reduction factor ($2^{(2n-f)}$/number of inheritance vector classes)	1	3.2	5.82	18.29	42.67

ASP = Affected sib pair; AST = affected sib triplet; ASQ = affected sib quadruplet; ASQui = affected sib quintet; ASS = affected sib sextet.

SpeedUp

The initial effort of the algebraic algorithm to identify the inheritance vector classes of all pedigrees is high, but the ensuing calculation of LOD scores assuming a large number of disease models is sped up considerably, especially when a dataset is comprised of pedigrees of only a few types. For example, in a dataset of 1,000 ASTs, the disease-locus-likelihood contributions of the 5 inheritance vector classes, given a certain disease model, have to be calculated only once for the whole dataset rather than 1,000 times.

The Peeling Algorithm

In the original version of GHM, the calculation of the parametric disease-locus likelihood is done separately for each inheritance vector by applying the Elston-Stewart algorithm, i.e. peeling nuclear

Table 2. Allele-sharing classes for discordant scenarios

	DSP	DSQ	DML	D3G
Number of inheritance vectors ($2^{(2n-f)}$)	4	64	64	128
Inheritance vector classes with imprinting taken into account	4	28	64	80
Reduction factor ($2^{(2n-f)}$/number of inheritance vector classes)	1	2.29	1	1.6

DSP = Discordant sib pair; DSQ = discordant sib quadruplet; DML = discordant marriage loop; D3G = discordant three-generation pedigree.

Table 3. Overview of scenarios for run-time assessment

Dataset No.	1	2	3	4	5	6	7	8
Pedigree type	ASPs	ASTs	ASQs	ASQuis	ASSs	equal mixture of 1–5	D3Gs	discordant mixture

For each dataset, 100 pedigrees were simulated using SLINK [51–53] for the genotype data at the disease locus and the SLINK utility program SUP [51, 54] for the marker genotypes.

Disease model $\{f_0, f_1, f_2\} = \{0.01, 0.1, 0.2\}$; p = 0.05.

Disease locus halfway between marker No. 50 and 51.

We used the following analysis options: 'imprinting on', 'algebraic calculation on/off', 'dimensions 5', 'saved models 0/5,000', 'number of replicates 1,000', 'maximization dense', 'penetrance restriction off', 'allfreq restriction off', 'analysis LOD', 'modcalc single', and 'calculate p value'.

families of the pedigree, to the disease locus. For the final remaining nuclear family of the pedigree or if the pedigree consists of only a single nuclear family, e.g. an ASP, a brute force calculation is employed. This calculation is done numerically and separately for each inheritance vector and for each assumed set of trait-model parameters. The LOD score of the currently analyzed family is stored, and the calculation continues with the next pedigree in the dataset. With the GHM software, many disease models are evaluated in a single program run during MOD-score analysis by repeating this step of the likelihood calculation. Our new algebraic procedure for calculating the disease-locus likelihood completely replaces the peeling algorithm, and it is applicable without additional modifications in case of inbreeding and marriage loops. It therefore significantly decreases the run time of a linkage analysis for any type of pedigree.

Maximization Options of GHM

The maximization routine of GHM first evaluates a set of predefined models. The user can choose between predefined grids with different densities. Moreover, the maximization can either be performed separately for each tested locus ('modcalc single' option) or jointly for the entire genetic region ('modcalc global' option). With modcalc single, calculation time can be saved by storing the trait-model-specific arrays of the disease-locus likelihood, which are needed for every considered genetic position. This option ('saved models') is especially useful when simulations are performed to obtain p values, which is already available with the previous version of GHM ('calculate p value' option [10]).

Simulations

To demonstrate the performance of our new method, we simulated datasets and compared the analysis run times of the algebraic algorithm to those of the peeling algorithm, which is employed by the original version that performs numeric calculation. Datasets either consisted of a single pedigree type, i.e. affected sibships with 2–6 siblings or three-generation pedigrees including unaffected pedigree members (discordant pedigrees), or mixtures of affected sibships. The speedup of the algebraic algorithm might be reduced by an increasing degree of discordance of the pedigrees, because this mostly leads to a larger number of inheritance vector classes as compared to their concordant counterparts (table 2). Therefore, we additionally considered an equal mixture of 4 discordant pedigree types: (a) discordant sib pairs, (b) discordant sib quadruplets, (c) discordant marriage loops (DML), and (d) discordant three-generation pedigrees (D3G). An overview of the simulated scenarios is given in table 3. Figure 3 depicts the pedigrees used for the discordant scenario including the one used in the D3G scenario (fig. 3d). Storing of arrays of the disease-locus likelihood, as already possible with the original GHM version (saved models option as mentioned above), was performed with the original algorithm (classic calculation mode). This was done to ensure a fair comparison to the classic calculation mode that makes use of run time-saving optimizations already implemented in the original GHM version. The saved models option was set to zero (no models saved) when using the algebraic algorithm (algebraic calculation mode), because it does not necessarily benefit from this option. It is of note that both our new

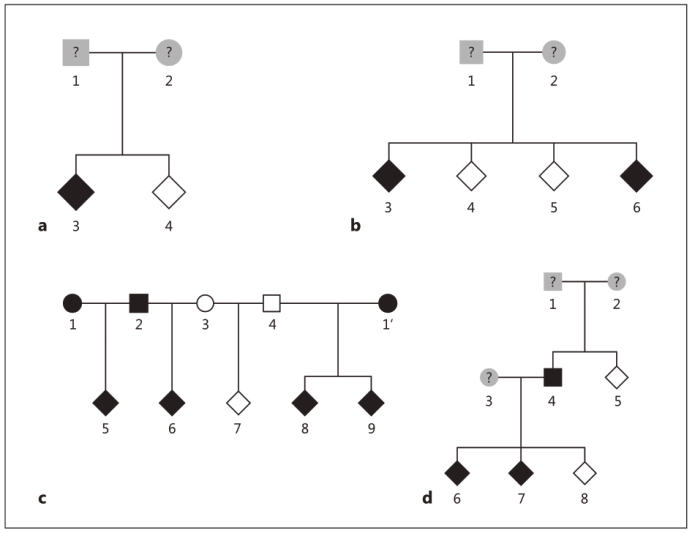


Fig. 3. Discordant pedigrees used in the simulations for run-time assessment. **a** Discordant sib pair; **b** discordant sib quadruplet; **c** discordant marriage loop; **d** discordant three-generation pedigree.

method and the saved models option need additional main memory. In case of the new method, this memory amount crucially depends on the size and phenotypic structure of the pedigrees, i.e. the number of inheritance vector classes across the whole dataset, whereas for the saved models option it depends on the number and size of the pedigrees. The peeling algorithm without the saved models option needs less memory albeit performing more floating point operations; it can still be used in case of insufficient main memory. Here we used a dense grid of disease models (option 'maximization dense'), because our new method should be especially useful when many disease models are evaluated, i.e. with a thorough maximization, which is likely to increase the power to map the disease gene under a complex mode of inheritance. In addition to the above-mentioned MOD-score analysis, p values were calculated (with the calculate p value option of GHM) by simulating 1,000 replicates generated under the null hypothesis of no linkage.

Run-Time Assessment

Run time was measured with the performance analysis tool *gprof* [46]. *gprof* measures the total amount of time spent executing each function of the program. Time due to system calls and waiting for CPU or I/O is not considered. Therefore, we additionally assessed the wall-clock time (WCT), which is the elapsed real time, i.e. the actual time taken from the start of the program run until the end. Because the WCT is obtained without any profiling steps, the program was run without any debugging options turned on. The speedup of our new method is obtained as follows:

$$\text{Speedup} = \frac{\text{run time with classic calculation mode}}{\text{run time with algebraic calculation mode}}.$$

All analyses were run on a single processor of the High Performance Computing – High Availability – Cluster (HPC-HA-Cluster) of the Helmholtz Zentrum München, equipped with IBM Intel Xeon X5690 6C, 3.46 GHz, 12 MB cache, 1,333 MHz 130 W processors in the compute nodes.

Results

The results of speedup due to the algebraic algorithm under the simulated scenarios for the analysis without calculating p values are shown in table 4, and those for the analysis with calculating p values are shown in table 5. Speedup is given based on run-time assessments measured by the performance analysis tool *gprof* as well as by measuring the WCT. Before looking at the speedups in detail, some technical aspects need to be considered prior to the interpretation of the results. In general, the *gprof* results reflect the speedup achieved by less time spent in the source code, which equals the number of instructions executed, but they do not include the time spent waiting for CPU and memory. Concerning GHM, the percentage of run time due to time waiting for CPU and memory increases with a larger number of scoring-function arrays saved in memory (saved models option) in case of the classic calculation mode, or with a larger number of inheritance vectors that must be considered when identify-

Table 4. Results of the run-time assessment without calculating p values, averaged over 3 program runs

	Run time, s				Speedup	
	WCT		*gprof*		WCT	*gprof*
	classic	algebraic	classic	algebraic		
Pedigree type						
ASPs	34.65	17.90	28.84	8.73	1.94	3.30
ASTs	95.75	23.76	73.97	15.13	4.03	4.89
ASQs	309.10	43.67	281.38	31.61	7.08	8.90
ASQuis	884.86	99.33	871.10	91.02	8.91	9.57
ASSs	4,523.33	392.67	3,372.38	378.24	11.52	8.92
Affected mixture	1,010.96	123.47	911.68	106.01	8.19	8.60
D3Gs	400.30	83.00	277.92	71.66	4.83	3.88
Discordant mixture	780.99	105.24	758.98	94.09	7.42	8.07

Classic = MOD-score analysis using the original GHM version; algebraic = MOD-score analysis using our new algebraic algorithm; *gprof* = execution time as measured by the profiling software *gprof*; ASQs = affected sib quadruplets; ASQuis = affected sib quintets; Affected mixture = mixture of 20 ASPs, ASTs, ASQs, ASQuis, and ASSs each; D3Gs = sample depicted in figure 3d; Discordant mixture = mixture of discordant pedigrees, 25 of each sort depicted in figure 3.

Table 5. Results of the run-time with calculating p values, averaged over 3 program runs

	Run time, h				Speedup	
	WCT		*gprof*		WCT	*gprof*
	classic	algebraic	classic	algebraic		
Pedigree type						
ASPs	9.28	5.49	5.18	2.22	1.69	2.33
ASTs	18.74	6.92	9.71	3.22	2.71	3.02
ASQs	62.85	11.68	17.96	4.95	5.38	3.63
ASQuis	243.48	31.85	26.98	9.01	7.64	2.99
ASSs	1,055.51	101.92	34.49	14.87	10.36	2.32
Affected mixture	278.97	33.98	28.81	9.56	8.21	3.01
D3Gs	177.80	20.39	36.84	8.29	8.72	4.44
Discordant mixture	294.83	29.36	29.20	10.69	10.04	2.73

See legend of table 4 for explanations.

ing inheritance vector classes in case of the algebraic calculation mode. For the latter, this is due to an increasing number of CPU memory cache misses caused by many crisscross copying processes of disease-locus-likelihood contributions of inheritance vectors of a given class into the corresponding memory cells of the scoring-function array. This copying process to complete the scoring-function has to be done for each inheritance vector, because $P_{complete}$, which will be multiplied with the scoring function, can be different for inheritance vectors of the same class. Hence, a larger number of inheritance vectors leads to more such copying processes, irrespective of the reduction factors as calculated in tables 1 and 2. When p values are calculated, this effect becomes more pronounced, as scoring-function arrays must be filled in this manner for every simulated replicate. In addition, it is of note that the results for the analyses without calculating p values are subject to a larger variance than those with calculating p values, because the analyses without calculating p values took only seconds to a few minutes to complete. With regard to the results in table 4 for the analyses without calculating p values, time waiting for CPU and memory was

almost negligible. This is due to the fact that, in addition to time spent for the initial preparation of the dataset, time was predominantly spent for the initial identification of inheritance vector classes in case of the algebraic calculation mode or the initial numeric calculation of scoring-function arrays used for model saving in case of the classic calculation mode with the saved models option. Hence, the *gprof* speedups of the scenarios without calculating p values in table 4 were similar to their corresponding speedups calculated from the WCT. On the contrary, the *gprof* speedups of the scenarios with calculating p values in table 5 were quite constant over varying pedigree types due to a larger percentage of function calls invoked by the calculate p value option, which remained unchanged in the new GHM version. In addition, most of the computing time as measured by the WCT was spent waiting for CPU and memory (see explanation above). As the WCT is more relevant for users, since it is the actual time they have to wait for results, we concentrate our discussion of speedup on the WCT. As can be seen in table 4, the speedup for the analysis without calculating p values ranged from 1.94 for ASPs to 11.52 for affected sib sextets (ASSs). These speedups turned out to be roughly proportional to the reduction factors as calculated in table 1. The speedup for the mixture of nuclear families (8.19) was approximately the average of the individual speedups for each pedigree type. The speedups of the D3G and the discordant scenarios were 4.83 and 7.42, respectively, which are higher than would have been expected from the reduction factors in table 2. The fact that the increased computational effort of the peeling algorithm to calculate the disease-locus likelihoods of the D3G and DML pedigrees is avoided with our new algorithmic approach might be responsible for that. When p values were calculated, the speedups for the scenarios of nuclear families ranged from 1.69 for ASPs to 10.36 for ASSs (table 5), as was expected from the reduction factors calculated in table 1. Even though the classic calculation mode took advantage of model saving, whose effect should be more pronounced when simulating replicates to calculate p values, the speedups from table 5 for nuclear families were similar to those from table 4. The speedup for the mixture of nuclear families was 8.21, which was again roughly the average of the individual speedups for each pedigree type. The speedups of the D3G and the discordant scenarios were 8.72 and 10.04, respectively. Here, the speedups were higher compared to the results without calculating p values given in table 4. This is due to the fact that the percentage of time needed for peeling of the D3G and DML pedigrees with the classic calculation mode is even more pronounced when p values are calculated, because time due to initial calculations, i.e. the identification of inheritance vector classes for the algebraic calculation mode and the initial calculation of scoring-function arrays for the classic calculation mode, was negligible.

Discussion

The calculation of the disease-locus likelihood in linkage analysis is a complex task, because data on the observed genetic markers are often incomplete. This leads to a large number of possible disease-locus genotypes that must be considered in the likelihood. MOD-score analysis is a promising route to the genetic dissection of complex traits in the context of family studies. Although time-consuming, the evaluation of many disease models during a MOD-score analysis is essential, because it is thus likely to increase the power to map genes that act under a complex mode of inheritance, compared to a simple parametric (LOD-score) or nonparametric (NPL-score) analysis.

Our algebraic algorithm is inspired by the identity of the allele-sharing-based nonparametric likelihood and the parametric likelihood in the test for linkage. It is based on the concept of inheritance vectors. These are collapsed into inheritance vector classes, which turn out to be the distinct allele-sharing classes in the nonparametric context. In the Appendix section, we theoretically derive the allele-sharing classes for the example of an AST when an imprinting model is considered. This tedious way of identifying allele-sharing classes could principally be done for any type of pedigree considering affected as well as unaffected pedigree members in order to construct an allele-sharing-based test for linkage (see also Strauch [22]). Due to the above-mentioned identity, however, it is straightforward to express the allele-sharing probabilities as functions of the trait-model parameters f_0, f_1, f_2, and p, and to perform a MOD-score analysis, i.e. the parametric equivalent of the nonparametric test. The algebraic algorithm can thus be considered as a unified approach of parametric and nonparametric linkage methods. Previous work has shown that the MOD-score approach can outperform other linkage methods in terms of power [9]. One of the reasons for this finding is the fact that the performance of LOD scores crucially depends on the specification of the correct trait model, which is generally unknown when analyzing complex traits. This problem is circumvented by the MOD score which, in contrast to the simple LOD score, is maximized not only over the recom-

bination fraction but also over trait-model parameters. However, the calculation of the disease-locus likelihood has to be done anew for every tested set of trait-model parameters, and it is the most time-consuming step in a MOD-score analysis. As a further complication, MOD scores are inflated when compared to LOD scores, and simulations to calculate p values have to be performed. Both aspects, extensive model testing and simulations to calculate p values, pose a challenge in regard to computation time and memory demands.

In this paper, we have presented a new algebraic algorithm that considerably reduces the run time of a MOD-score analysis. By storing unique IVDLGCs in a database common to all pedigrees in a dataset, the number of floating point operations and the memory demand of our new method are kept minimal, and similarities of family trees in terms of disease-locus-likelihood contributions can be exploited across the whole dataset. This is possible because the disease locus is treated separately from the marker loci when using a linkage analysis program such as GHM [10–13] that is based on the Lander-Green algorithm [25]. The speedup of a linkage analysis with GHM due to the algebraic algorithm depends on the number of different pedigree types, the complexity of the pedigrees, which is expressed by the number of inheritance vectors and classes, the number of replicates used to calculate p values, and the number of models saved in memory (saved models option) when running GHM in the classic calculation mode. For datasets consisting of only a single type of nuclear families, the speedup increased with the number of affected siblings and reached a factor >10 for ASSs in our analyses (tables 4, 5). Even in the case of ASPs, we achieved speedups by a factor of more than 1.5 (tables 4, 5). When using an equal mixture of nuclear families with different numbers of affected offspring, the speedups turned out to be the approximate average of the speedups of the individual nuclear family scenarios (tables 4, 5). In the D3G and the discordant scenarios, i.e. those scenarios with a larger degree of complexity of the pedigrees and a higher computational burden due to peeling and loop breaking for the classic calculation mode of GHM, the speedups increased from the analysis without calculating p values to those with calculating p values from 4.83 to a factor >8 for the D3G scenario, and from 7.42 to a factor >10 for the discordant scenario. The results thus clearly show that our new algorithm can substantially reduce the run time of a MOD-score analysis with GHM.

In the past, linkage analysis proved to be a valuable tool for identifying regions of the genome that harbor variants responsible for both Mendelian and complex diseases [2]. However, sequencing a rather large genetic region represented by the linkage signal to determine the causal variant was not feasible at that time. Nowadays, employing next-generation sequencing techniques allows for the identification of rare causal variants of putative complex-disease genes by combining an initial step of linkage analysis followed by fine mapping with association analysis. A major advantage of linkage methods as compared to methods in association analysis is that information across families can be combined, such that evidence for a causal role of a locus can accumulate even if different variants segregate at that locus in different families, which is known as allelic heterogeneity [47]. However, locus heterogeneity and/or penetrance heterogeneity, i.e. several allelic variants exist at the same locus each with different penetrances, can reduce the power of linkage analysis to map the disease gene. This problem can be diminished using large pedigrees, which can each be more homogeneous with respect to genetic variation than unrelated individuals or a sample of many small pedigrees [48, 49]. Admittedly, the GENEHUNTER software was originally designed for the analysis of small to moderately sized pedigrees ($2n - f \leq 20$ with n non-founders and f founders in a pedigree). Such pedigrees are easier to collect for diseases characterized by late onset, low penetrance, and diagnostic uncertainty. They are also more likely to reflect the genetic etiology of the disease in the general population [35]. The loss of power due to the uncertainty in penetrance values at the disease locus can be reduced by a maximization of the disease-locus likelihood over the trait-model parameters f_0, f_1, f_2, and p as it is done in a MOD-score analysis. Further robustness can be obtained by performing an affecteds-only analysis through recoding unaffected individuals as having an unknown phenotype. If the penetrance is low, little information is lost by ignoring the phenotype of unaffected pedigree members. The power of an affecteds-only MOD-score analysis can hence be higher, because the MOD-score distribution has fewer degrees of freedom as compared to the MOD score in an analysis that uses the phenotype of unaffected pedigree members. Even if pedigrees show locus and/or penetrance heterogeneity, it is likely that modest evidence for linkage can indeed narrow down the genetic region harboring the disease gene and can hence be used as a filter to focus on a more detailed association analysis of the variants in the region. In addition, using large samples of small pedigrees allows for the identification of hitherto unidentified genetic variants as risk factors for complex diseases (see de Visser et al. [50] for an example with

ASPs). Therefore, while linkage analysis of rare variants segregating in large pedigrees has proven to be a powerful approach, the analysis of smaller pedigrees can also be a promising route to discover genetic loci responsible for complex traits by the use of whole-exome or whole-genome sequence data. Irrespective of the assumed underlying genetic architecture of a given collection of small pedigrees, e.g. a large number of small-effect common variants, a large number of large-effect rare variants, or a mixture of both, GHM is well suited for the analysis of such data.

Extensive model testing, simulations to calculate p values, and the consideration of many genetic markers in a MOD-score analysis are indispensable to successfully map complex-disease genes in the context of family studies. Our new algebraic algorithm paves the way to an exceedingly efficient MOD-score analysis, because the evaluation of many sets of trait-model parameters and simulations to calculate p values are now feasible within a reasonable amount of time. Assuming, for example, an average speedup of 6.84 calculated from table 5, a geneticist doing a linkage study with MOD scores including simulations to determine p values can obtain results within a day instead of waiting a whole week for the analysis to finish. This further pushes ahead the maximum size of pedigrees that can still be analyzed.

GENEHUNTER-MODSCORE is thus a promising tool to identify rare causal variants segregating within families using next-generation-sequencing data. The algebraic algorithm is implemented in a new version of GHM that can be obtained for free from the following website: www.helmholtz-muenchen.de/ige/service/software-download/index.html.

Acknowledgements

This work was supported by grants Str643/4-1 and Str643/6-1 of the Deutsche Forschungsgemeinschaft (German Research Foundation). We thank Clemens Baumbach for his help with the figures. In addition, we thank the reviewers for their thoughtful comments, which have helped to improve the paper.

Appendix

Calculation of Allele-Sharing Classes for an AST Taking Imprinting into Account (see also Knapp [23] for the Formulation without Imprinting)

We are interested in the IBD sharing probability distribution of an AST at a diallelic disease locus with susceptibility allele D, normal allele d, and allele frequencies $p = P(D)$, $q = P(d) = 1 - p$. Taking the parental origin of the alleles into account, 5 IBD configurations can be distinguished. These IBD configurations are identical to the inheritance vector classes. Table 1A presents the Mendelian probability for each IBD configuration and a representative sharing among the 3 sibs. Let w_i^D ($i = 0, 1, 2^{pat}, 2^{mat}$, and 3) denote the probability of the i-th configuration at the disease locus. Further, let D_p and D_m denote the paternal and maternal genotype at the disease locus. Let AST be the event that all 3 sibs are affected, and let IBD_i be the event that the sibs have IBD configuration i at the disease locus. For $k, l, m, n, \in \{D, d\}$, let $c_i^{(k,\,l,\,m,\,n)}$ denote the probability of the joint occurrence of AST and IBD_i^D, given that the paternal and maternal genotypes are (k, l) and (m, n). We hence get

$$\begin{aligned} c_i^{(k,\,l,\,m,\,n)} &= P\,(AST \cap IBD_i^D | D_{pat} = (k, l), D_{mat} = (m, n)) \\ &= P(AST | IBD_i^D, D_{pat} = (k, l), D_{mat} = (m, n)) \\ &\quad \cdot P(IBD_i^D | D_{pat} = (k, l), D_{mat} = (m, n)), \end{aligned}$$

where $P(IBD_i^D | D_{pat} = (k, l), D_{mat} = (m, n))$ reduces to the Mendelian probability of the i-th IBD configuration, i.e. $P(IBD_i^D)$.

With *first-bits* $\in \mathcal{G}$, $\mathcal{G} = \{00, 01, 10, 11\}$ denoting the first two bits of the inheritance vector, which correspond to the outcome of the two meioses leading to the first offspring, we obtain

$$\begin{aligned} &P(AST | IBD_i^D, D_{pat} = (k, l), D_{mat} = (m, n)) \\ &\quad = \Sigma_{\textit{first-bits} \,\in\, \mathcal{G}}\, P(AST, \textit{first-bits} | IBD_i^D, D_{pat} = (k, l), D_{mat} = (m, n)) \\ &\quad = \Sigma_{\textit{first-bits} \,\in\, \mathcal{G}}\, P(AST | \textit{first-bits}, IBD_i^D, D_{pat} = (k, l), D_{mat} = (m, n)) \\ &\quad \cdot P\,(\textit{first-bits} | IBD_i^D, D_{pat} = (k, l), D_{mat} = (m, n)), \end{aligned}$$

where $P\,(\textit{first-bits} | IBD_i^D, D_{pat} = (k, l), D_{mat} = (m, n)) = 1/4$ for all *first-bits* $\in \{00, 01, 10, 11\}$.

Thus, we can write for $c_i^{(k,\,l,\,m,\,n)}$

$$c_i^{(k,\,l,\,m,\,n)} = 1/4\; P(IBD_i^D)\; \Sigma_{\textit{first-bits} \,\in G}\, P(AST | \textit{first-bits}, IBD_i^D, D_{pat} = (k, l), D_{mat} = (m, n)).$$

Then with $\mathcal{J} = \{D, d\}^4$, it follows

$$\sum_{(k,\,l,\,m,\,n) \in \mathcal{J}} c_i^{(k,\,l,\,m,\,n)}\, P\left(D_{pat} = (k, l), D_{mat} = (m, n)\right) = P\left(AST | IBD_i^D\right) \cdot P\left(IBD_i^D\right) = P\left(AST \cap IBD_i^D\right)$$

and further

$$w_i^D = P\left(IBD_i | AST\right) = \frac{\sum_{(k,\,l,\,m,\,n) \in \mathcal{J}} c_i^{(k,\,l,\,m,\,n)} P\left(D_{pat} = (k, l), D_{mat} = (m, n)\right)}{P(AST)}.$$

For the 5 inheritance vector classes in the context of ASTs we obtain:

$$\begin{aligned} c_3^{(k,\,l,\,m,\,n)} &= 1/64\; (f^3_{km} + f^3_{kn} + f^3_{lm} + f^3_{ln}) \\ c_{2,pat}^{(k,\,l,\,m,\,n)} &= 3/64\; (f_{km}f_{kn}(f_{km} + f_{kn}) + f_{lm}f_{ln}(f_{lm} + f_{ln})) \\ c_{2,mat}^{(k,\,l,\,m,\,n)} &= 3/64\; (f_{km}f_{lm}(f_{km} + f_{lm}) + f_{kn}f_{ln}(f_{kn} + f_{ln})) \\ c_1^{(k,\,l,\,m,\,n)} &= 3/64\; (f_{km}f_{ln}(f_{km} + f_{ln}) + f_{kn}f_{lm}(f_{kn} + f_{lm})) \\ c_0^{(k,\,l,\,m,\,n)} &= 3/32\; (f_{km}f_{kn}(f_{lm} + f_{ln}) + f_{lm}f_{ln}(f_{km} + f_{kn})) \end{aligned}$$

$$w_3^D = \frac{1}{16P(AST)}\left(p^2 f_2^3 + pqf_{1,pat}^3 + pqf_{1,mat}^3 + q^2 f_0^3\right)$$

$$w_{2,pat}^D = \frac{3}{16P(AST)}\begin{pmatrix} p^3 f_2^3 + p^2 q\left(f_{1,pat} f_{1,mat}^2 + f_2^2 f_{1,pat} + f_2 f_{1,pat}^2\right) \\ + pq^2\left(f_{1,mat}^3 + f_{1,mat}^2 f_0 + f_{1,mat} f_0^2\right) \end{pmatrix}$$

$$w_{2,mat}^D = \frac{3}{16P(AST)}\begin{pmatrix} p^3 f_2^3 + p^2 q\left(f_{1,pat}^2 f_{1,mat} + f_2^2 f_{1,mat} + f_2 f_{1,mat}^2\right) \\ + pq^2\left(f_{1,pat}^3 + f_{1,pat}^2 f_0 + f_{1,pat} f_0^2\right) \end{pmatrix}$$

$$w_1^D = \frac{3}{16P(AST)}\begin{pmatrix} \left(p^2 f_2^2 + pqf_{1,pat}^2 + pqf_{1,mat}^2 + q^2 f_0^2\right) \\ \cdot\left(p^2 f_2 + pqf_{1,pat} + pqf_{1,mat} + q^2 f_0\right) \end{pmatrix}$$

$$w_0^D = \frac{3}{8P(AST)}\begin{pmatrix} p^4 f_2^3 + p^3 qf_2\left(f_2 f_{1,pat} + f_{1,pat}^2 + f_2 f_{1,mat} + f_{1,mat}^2\right) \\ + p^2 q^2\begin{pmatrix} f_{1,pat}^3 + f_{1,mat}^3 + f_2 f_{1,pat} f_{1,mat} + f_2 f_{1,pat} f_0 \\ + f_2 f_{1,mat} f_0 + f_{1,pat} f_{1,mat} f_0 \end{pmatrix} \\ + pq^3 f_0\left(f_{1,pat}^2 + f_{1,pat} f_0 + f_{1,mat}^2 + f_{1,mat} f_0 + q^4 f_0^3\right) \end{pmatrix}.$$

Table 1A. IBD configurations for three affected siblings A, B, and C (adapted from Knapp [23])

IBD configuration/ inheritance vector class i	Alleles shared IBD by			Mendelian probability
	AB	AC	BC	
3	2	2	2	1/16
2^{pat}	2	1^{pat}	1^{pat}	3/16
2^{mat}	2	1^{mat}	1^{mat}	3/16
1	2	0	0	3/16
0	1^{pat}	0	1^{mat}	3/8

For each IBD configuration, i.e. inheritance vector class, i, the Mendelian probability and a representative sharing among the 3 siblings are given. Note that the 3 siblings A, B, and C cannot be distinguished, such that e.g. siblings A and C could be flipped, which reduces the number of inheritance vector classes. Hence, with 16 inheritance vectors for an AST, the Mendelian probability of e.g. inheritance vector class $i = 1$ is 3/16, because the sharing of 2 alleles IBD can take place either between A and B, A and C, or B and C, which does not have to be distinguished.

Table 2A. Mating types and conditional probabilities ci (adapted from Knapp [23])

No.	Parental mating type (pat × mat)	Probability of mating type	c_3	$c_{2,pat}$	$c_{2,mat}$	c_1	c_0	$\sum_i c_i$
1	DD × DD	p^4	$1/16\, f^3_2$	$3/16\, f^3_2$	$3/16\, f^3_2$	$3/16\, f^3_2$	$3/8\, f^3_2$	f^3_2
2	DD × Dd	$2p^3q$	$1/32\,(f^3_2 + f^3_{1,pat})$	$3/32\,(f^2_2 f_{1,pat} + f_2 f^2_{1,pat})$	$3/32\,(f^3_2 + f^3_{1,pat})$	$3/32\,(f^2_2 f_{1,pat} + f_2 f^2_{1,pat})$	$3/16\,(f^2_2 f_{1,pat} + f_2 f^2_{1,pat})$	$1/8\,(f_2 + f_{1,pat})^3$
3	Dd × DD	$2p^3q$	$1/32\,(f^3_2 + f^3_{1,mat})$	$3/32\,(f^3_2 + f^3_{1,mat})$	$3/32\,(f^2_2 f_{1,mat} + f_2 f^2_{1,mat})$	$3/32\,(f^2_2 f_{1,mat} + f_2 f^2_{1,mat})$	$3/16\,(f^2_2 f_{1,mat} + f_2 f^2_{1,mat})$	$1/8\,(f_2 + f_{1,mat})^3$
4	DD × dd	p^2q^2	$1/16\, f^3_{1,pat}$	$3/16\, f^3_{1,pat}$	$3/16\, f^3_{1,pat}$	$3/16\, f^3_{1,pat}$	$3/8\, f^3_{1,pat}$	$f^3_{1,pat}$
5	dd × DD	p^2q^2	$1/16\, f^3_{1,mat}$	$3/16\, f^3_{1,mat}$	$3/16\, f^3_{1,mat}$	$3/16\, f^3_{1,mat}$	$3/8\, f^3_{1,mat}$	$f^3_{1,mat}$
6	Dd × Dd	$4p^2q^2$	$1/64\,(f^3_2 + f^3_{1,pat} + f^3_{1,mat} + f^3_0)$	$3/64\,(f^2_2 f_{1,pat} + f_2 f^2_{1,pat} + f^2_{1,mat} f_0 + f_{1,mat} f^2_0)$	$3/64\,(f^2_2 f_{1,mat} + f_2 f^2_{1,mat} + f^2_{1,pat} f_0 + f_{1,pat} f^2_0)$	$3/64\,(f^2_2 f_0 + f_2 f^2_0 + f^2_{1,pat} f_{1,mat} + f_{1,pat} f^2_{1,mat})$	$3/32\,(f_2 f_{1,pat} f_{1,mat} + f_2 f_{1,pat} f_0 + f_2 f_{1,mat} f_0, + f_{1,pat} f_{1,mat} f_0)$	$1/64\,(f_2 + f_{1,pat} + f_{1,mat} + f_0)^3$
7	Dd × dd	$2pq^3$	$1/32\,(f^3_{1,pat} + f^3_0)$	$3/32\,(f^3_{1,pat} + f^3_0)$	$3/32\,(f^2_{1,pat} f_0 + f_{1,pat} f^2_0)$	$3/32\,(f^2_{1,pat} f_0 + f_{1,pat} f^2_0)$	$3/16\,(f^2_{1,pat} f_0 + f_{1,pat} f^2_0)$	$1/8\,(f_{1,pat} + f_0)^3$
8	dd × Dd	$2pq^3$	$1/32\,(f^3_{1,mat} + f^3_0)$	$3/32\,(f^2_{1,mat} f_0 + f_{1,mat} f^2_0)$	$3/32\,(f^3_{1,mat} + f^3_0)$	$3/32\,(f^2_{1,mat} f_0 + f_{1,mat} f^2_0)$	$3/16\,(f^2_{1,mat} f_0 + f_{1,mat} f^2_0)$	$1/8\,(f_{1,mat} + f_0)^3$
9	dd × dd	q^4	$1/16\, f^3_0$	$3/16\, f^3_0$	$3/16\, f^3_0$	$1/16\, f^3_0$	$3/8\, f^3_0$	f^3_0

A diallelic disease locus with susceptibility allele D, normal allele d, and allele frequencies $p = P(D)$, $q = P(d) = 1 - p$ is assumed. If the order of alleles within a parent is ignored, 9 mating types $(k, l, m, n) \in J$, with $J = \{D, d\}^4$ have to be distinguished. The mating type probabilities are given under the assumption of Hardy-Weinberg equilibrium at the disease locus. $c_i^{(k,\,l,\,m,\,n)}$ denotes the probability of the joint occurrence of 3 affected sibs that have IBD configuration i at the disease locus, given that the paternal and maternal genotypes are (k, l) and (m, n). (f_0, $f_{1,pat}$, $f_{1,mat}$, and f_2) are the penetrances with f_i denoting the probability that an individual with i copies of the disease allele develops the disease. For the heterozygous individuals, separate penetrances for paternal and maternal transmission of the disease allele are distinguished to take imprinting into account.

References

1 Mohr J: A Study of Linkage in Man. Copenhagen, Munksgaards Forlag, 1954.
2 Bailey-Wilson JE, Wilson AF: Linkage analysis in the next-generation sequencing era. Hum Hered 2011;72:228–236.
3 Bowden DW: Will family studies return to prominence in human genetics and genomics? Rare variants and linkage analysis of complex traits. Genes Genom 2011;33:1–8.
4 Wilson AF, Ziegler A: Lessons learned from Genetic Analysis Workshop 17: transitioning from genome-wide association studies to whole-genome statistical genetic analysis. Genet Epidemiol 2011;35(suppl 1):S107–S114.
5 Morton NE: Sequential tests for the detection of linkage. Am J Hum Genet 1955;7:277–318.
6 Risch N: Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 1984;36:363–386.
7 Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J: Effects of misspecifying genetic parameters in lod score analysis. Biometrics 1986;42:393–399.
8 Risch N, Giuffra L: Model misspecification and multipoint linkage analysis. Hum Hered 1992;42:77–92.
9 Flaquer A, Strauch K: A comparison of different linkage statistics in small to moderate sized pedigrees with complex diseases. BMC Res Notes 2012;5:411.
10 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;32:73–83.
11 Dietter J, Mattheisen M, Fürst R, Rüschendorf F, Wienker TF, Strauch K: Linkage analysis using sex-specific recombination fractions with GENEHUNTER-MODSCORE. Bioinformatics 2007;23:64–70.
12 Strauch K, Fürst R, Rüschendorf F, Windemuth C, Dietter J, Flaquer A, Baur MP, Wienker TF: Linkage analysis of alcohol dependence using MOD scores. BMC Genet 2005;6(suppl 1):S162.
13 Strauch K: Parametric linkage analysis with automatic optimization of the disease model parameters. Am J Hum Genet 2003;73(suppl 1):A2624.
14 Flaquer A, Baumbach C, Piñero E, García Algas F, de la Fuente Sanchez MA, Rosell J, Toquero J, Alonso-Pulpon L, Garcia-Pavia P, Strauch K, Heine-Suñer D: Genome-wide linkage analysis of congenital heart defects using MOD score analysis identifies two novel loci. BMC Genet 2013;14:44.
15 Kruse LV, Nyegaard M, Christensen U, Møller-Larsen S, Haagerup A, Deleuran M, Hansen LG, Venø SK, Goossens D, Del-Favero J, Børglum AD: A genome-wide search for linkage to allergic rhinitis in Danish sib-pair families. Eur J Hum Genet 2012;20:965–972.
16 Christensen U, Møller-Larsen S, Nyegaard M, Haagerup A, Hedemand A, Brasch-Andersen C, Kruse TA, Corydon TJ, Deleuran M, Børglum AD: Linkage of atopic dermatitis to chromosomes 4q22, 3p24 and 3q21. Hum Genet 2009;126:549–557.
17 Schumacher J, Kaneva R, Jamra RA, et al: Genomewide scan and fine-mapping linkage studies in four European samples with bipolar affective disorder suggest a new susceptibility locus on chromosome 1p35–p36 and provides further evidence of loci on chromosome 4q31 and 6q24. Am J Hum Genet 2005;77:1102–1111.
18 Kurz T, Altmueller J, Strauch K, Rüschendorf F, Heinzmann A, Moffatt MF, Cookson WOCM, Inacio F, Nürnberg P, Stassen HH, Deichmann KA: A genome-wide screen on the genetics of atopy in a multiethnic European population reveals a major atopy locus on chromosome 3q21.3. Allergy 2005;60:192–199.
19 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;44:44–51.
20 Holmans P: Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 1993;52:362–374.
21 Suarez BK, Rice J, Reich T: The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 1978;42:87–94.
22 Strauch K: MOD-score analysis with simple pedigrees: an overview of likelihood-based linkage methods. Hum Hered 2007;64:192–202.
23 Knapp M: A note on linkage analysis with affected sib triplets. Hum Hered 2005;59:21–25.
24 Elston RC, Stewart J: A general model for the genetic analysis of pedigree data. Hum Hered 1971;21:523–542.
25 Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 1987;84:2363–2367.
26 Lathrop GM, Lalouel JM, White RL: Construction of human linkage maps: likelihood calculations for multilocus linkage analysis. Genet Epidemiol 1986;3:39–52.
27 Lathrop GM, Lalouel JM, Julier C, Ott J: Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 1984;81:3443–3446.
28 Lathrop GM, Lalouel JM: Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet 1984;36:460–465.
29 Schäffer AA, Gupta SK, Shriram K, Cottingham RW Jr: Avoiding recomputation in linkage analysis. Hum Hered 1994;44:225–237.
30 Cottingham RW Jr, Idury RM, Schäffer AA: Faster sequential genetic linkage computations. Am J Hum Genet 1993;53:252–263.
31 O'Connell JR: Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. Hum Hered 2001;51:226–240.
32 O'Connell JR, Weeks DE: The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 1995;11:402–408.
33 Gertz EM, Hiekkalinna T, Digabel SL, Audet C, Terwilliger JD, Schäffer AA: PSEUDOMARKER 2.0: efficient computation of likelihoods using NOMAD. BMC Bioinformatics 2014;15:47.
34 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;71:256–266.
35 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 1996;58:1347–1363.
36 Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsdottir A: Allegro version 2. Nat Genet. 2005;37:1015–1016.
37 Gudbjartsson DF, Jonasson K, Frigge ML, Kong A: Allegro, a new computer program for multipoint linkage analysis. Nat Genet 2000;25:12–13.
38 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;30:97–101.
39 Markianos K, Daly MJ, Kruglyak L: Efficient multipoint linkage analysis through reduction of inheritance space. Am J Hum Genet 2001;68:963–977.
40 Idury RM, Elston RC: A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. Hum Hered 1997;47:197–202.
41 Kruglyak L, Lander ES: Faster multipoint linkage analysis using Fourier transforms. J Comput Biol 1998;5:1–7.
42 Nyholt DR: GENEHUNTER: your 'one-stop shop' for statistical genetic analysis? Hum Hered 2002;53:2–7.

43 Whittemore AS, Halpern J: A class of tests for linkage using affected pedigree members. Biometrics 1994;50:118–127.

44 McPeek MS: Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet Epidemiol 1999;16:225–249.

45 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;66:1945–1957.

46 Graham SL, Kessler PB, McKusick MK: An execution profiler for modular programs. Software Pract Exper 1983;13:671–685.

47 Balding DJ: A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;7:781–791.

48 Wijsman EM: The role of large pedigrees in an era of high-throughput sequencing. Hum Genet 2012;131:1555–1563.

49 Terwilliger JD: On the resolution and feasibility of genome scanning approaches. Adv Genet 2001;42:351–391.

50 de Visser MCH, van Minkelen R, van Marion V, den Heijer M, Eikenboom J, Vos HL, Slagboom PE, Houwing-Duistermaat JJ, Rosendaal FR, Bertina RM: Genome-wide linkage scan in affected sibling pairs identifies novel susceptibility region for venous thromboembolism: Genetics In Familial Thrombosis study. J Thromb Haemost 2013;11:1474–1484.

51 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;71:126–134.

52 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;7:237–243.

53 Ott J: Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 1989;86:4175–4178.

54 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.

6. Paper II

Brugger M, Lutz M, Müller-Nurasyid M, Lichtner P, Slater EP, Matthäi E, Bartsch DK, Strauch K. Joint linkage and association analysis with GENEHUNTER-MODSCORE with an application to familial pancreatic cancer. Hum Hered. 2024;89(1):8—31.

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):

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

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, $f_{1,\text{pat}}$ and $f_{1,\text{mat}}$, 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:

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

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:

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

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: $\text{SNV}_1|\text{SNV}_2|\text{TL} \in \{0|0|0, 0|0|1, 0|1|0, 0|1|1, 1|0|0, 1|0|1, 1|1|0, 1|1|1\} = : \{h_0, \ldots, h_7\}$. 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:

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

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: $\text{SNV}_1|\text{SNV}_2|\text{SNV}_3|\text{TL} \in \{0|0|0|0, 0|0|0|1, 0|0|1|0, 0|0|1|1, 0|1|0|0, 0|1|0|1, 0|1|1|0, 0|1|1|1, 1|0|0|0, 1|0|0|1, 1|0|1|0, 1|0|1|1, 1|1|0|0, 1|1|0|1, 1|1|1|0, 1|1|1|1\} = : \{h_0, \ldots, h_{15}\}$. 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:

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

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 $p_{h_0},\ldots,p_{h_1}$ 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:

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

$\mathcal{F}$ represents the set of founders and $\mathcal{N}$ 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 $g\in\{0,$"1, pat","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:

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

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:

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

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 $g\in\{0,$"1, pat","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 $g\in\{0,$"1, pat","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}, $r = 0, \ldots, 2^m\text{-}1$ 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:

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

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

where *OFSG* denotes a particular set of ordered test SNV genotypes for all founders, and $p_{h_{OFSG_{k,1}}}^{EM(t)}$ and $p_{h_{OFSG_{k,2}}}^{EM(t)}$ are the haplotype frequencies for founder individual k of the paternally and maternally inherited haplotypes at the previous iteration t, respectively, with $OFSG_{k,1}, OFSG_{k,2} \in \{0, \ldots, 2^m - 1\}$. $z_{OFSG}^{h_r}$ 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:

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

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									
	$MAF_{1\ldots20}$			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

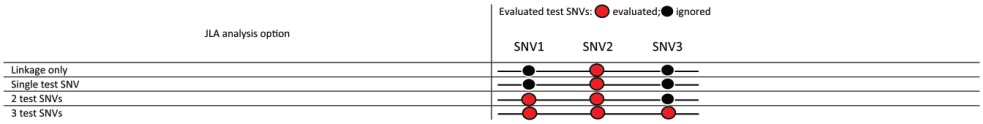
analysis, JLA analysis was performed using the left and the central test SNV (see also Table 4). The disease-allele frequency and penetrance restrictions were set to the default values ($p_m \leq 0.5; f_0 \leq f_{1,\text{pat}}, f_{1,\text{mat}} \leq f_2$). Imprinting analysis ($f_{1,\text{pat}} \neq f_{1,\text{mat}}$) 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 ($p_m \leq 0.5$; $f_0 \leq f_{1,\mathrm{pat}}, f_{1,\mathrm{mat}} \leq f_2$), and imprinting analysis ($f_{1,\mathrm{pat}} \neq f_{1,\mathrm{mat}}$) 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 phenocopy 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

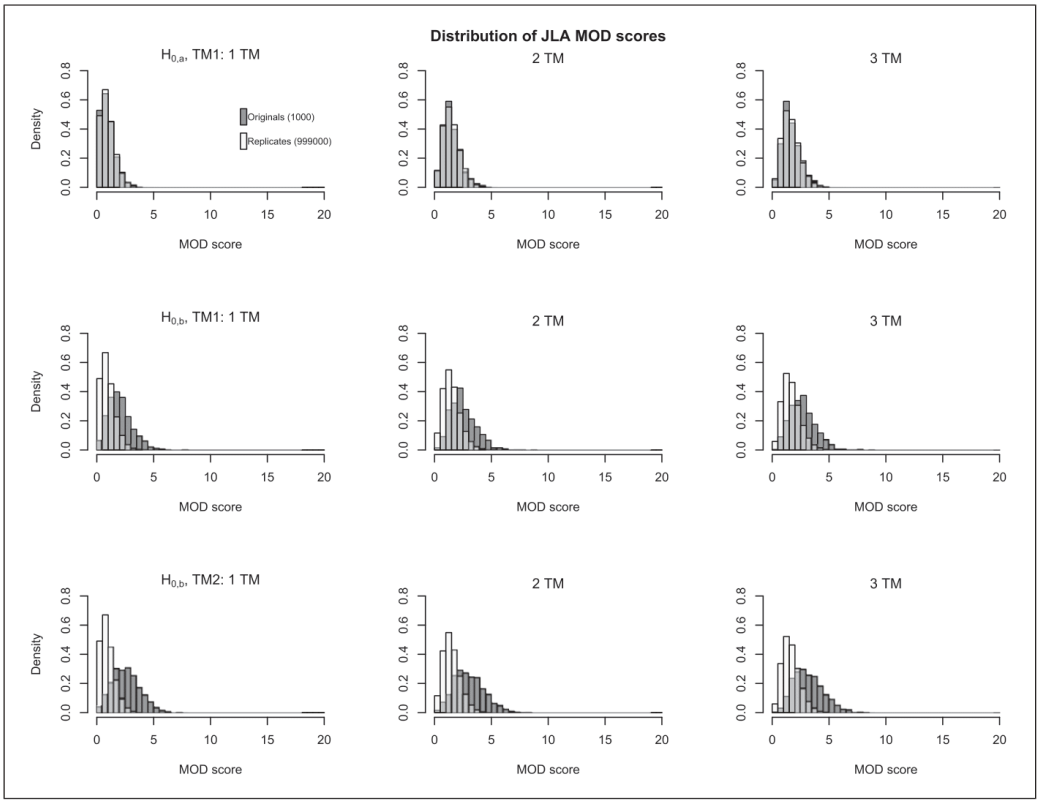


Fig. 1. Depiction of the empiric distributions of JLA-MOD scores for data simulated under the hypothesis of no linkage, no association (row 1, depiction only for trait model TM1) and linkage, no association (row 2 for TM1, row 3 for TM2). The bars of the JLA-MOD scores of the "original" simulated SLINK 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

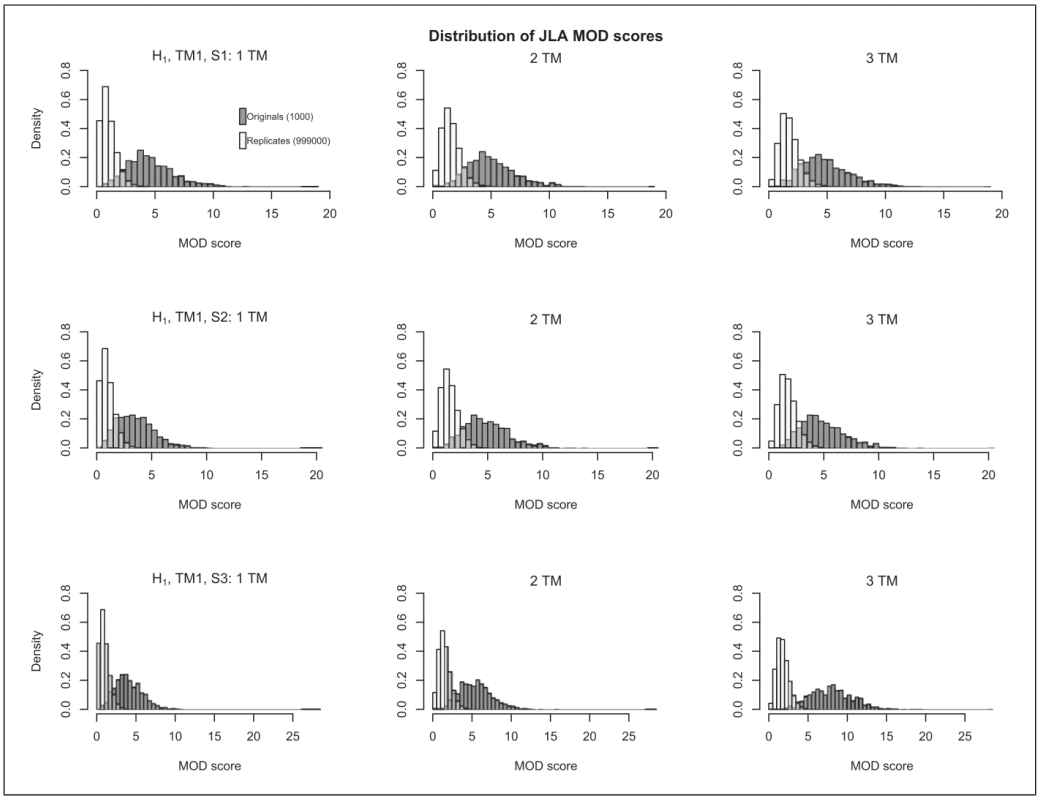


Fig. 2. Depiction of the empiric distributions of JLA-MOD scores for data simulated under the hypothesis of linkage and association for trait model TM1 and various LD patterns (row 1: S1; row 2: S2; row 3: S3). For more details, see Figure 1.

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

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

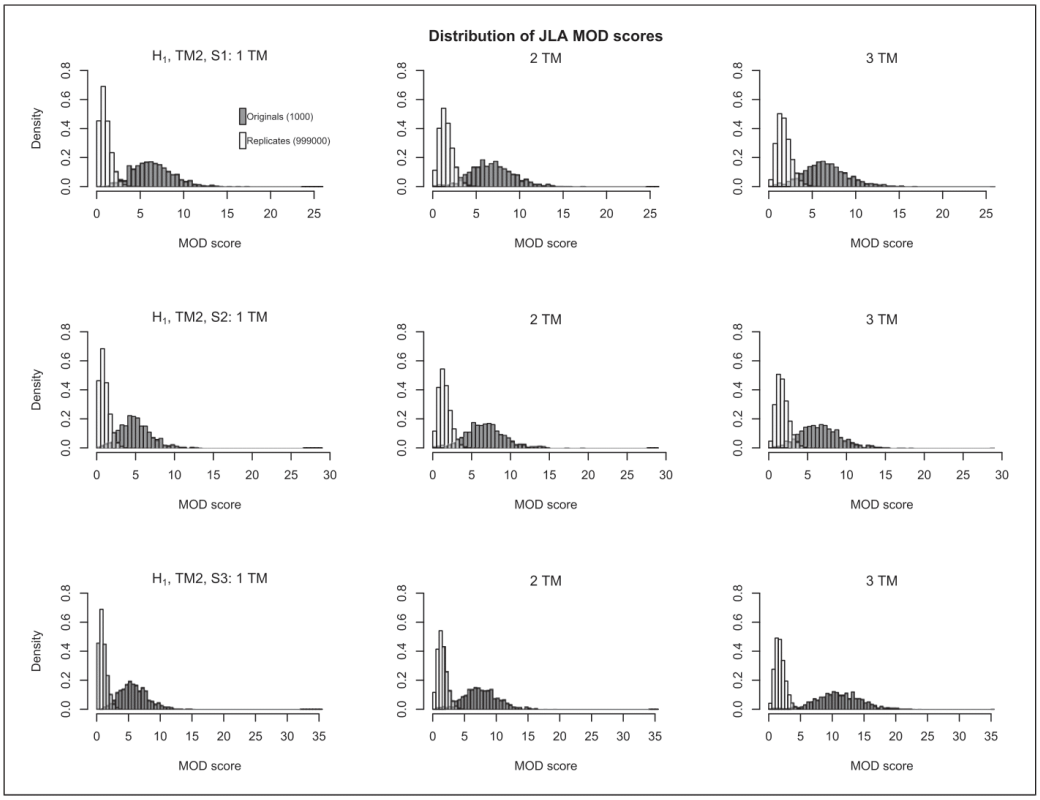


Fig. 3. Depiction of the empiric distributions of JLA-MOD scores for data simulated under the hypothesis of linkage and association for trait model TM2 and various LD patterns (row 1: S1; row 2: S2; row 3: S3). For more details, see Figure 1.

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.

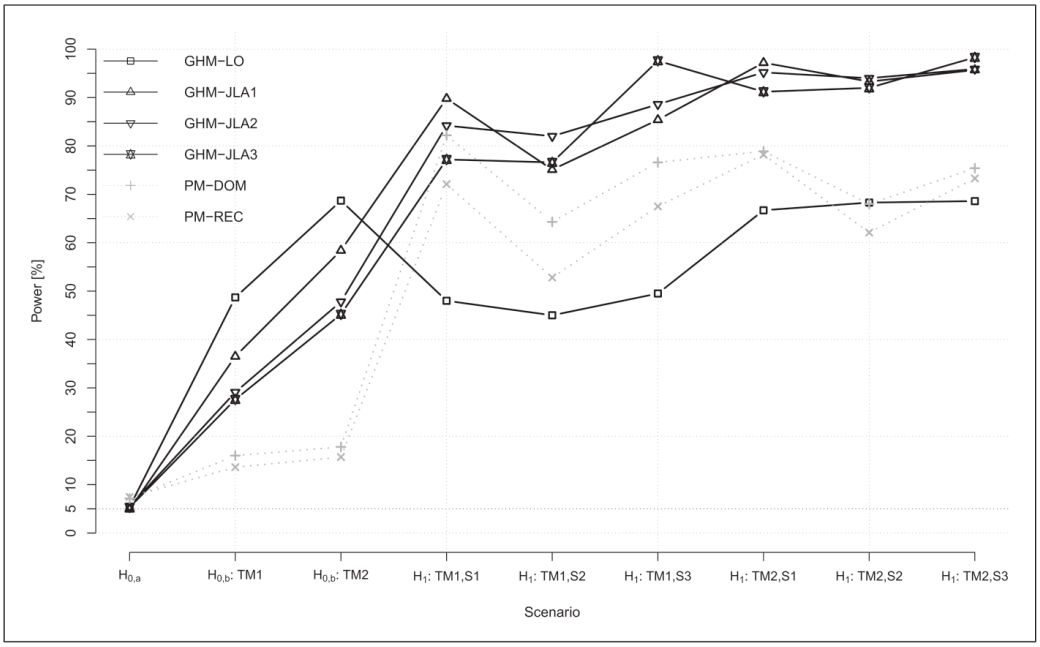


Fig. 4. Depiction of type I error and power values for the six investigated test statistics and analysis options: GHM-LO: GHM linkage-only MOD score; GHM-JLA1: GHM-JLA-MOD score using one test SNV for the analysis; GHM-JLA2: GHM-JLA-MOD score using two test SNVs for the analysis; GHM-JLA3: GHM-JLA-MOD 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.

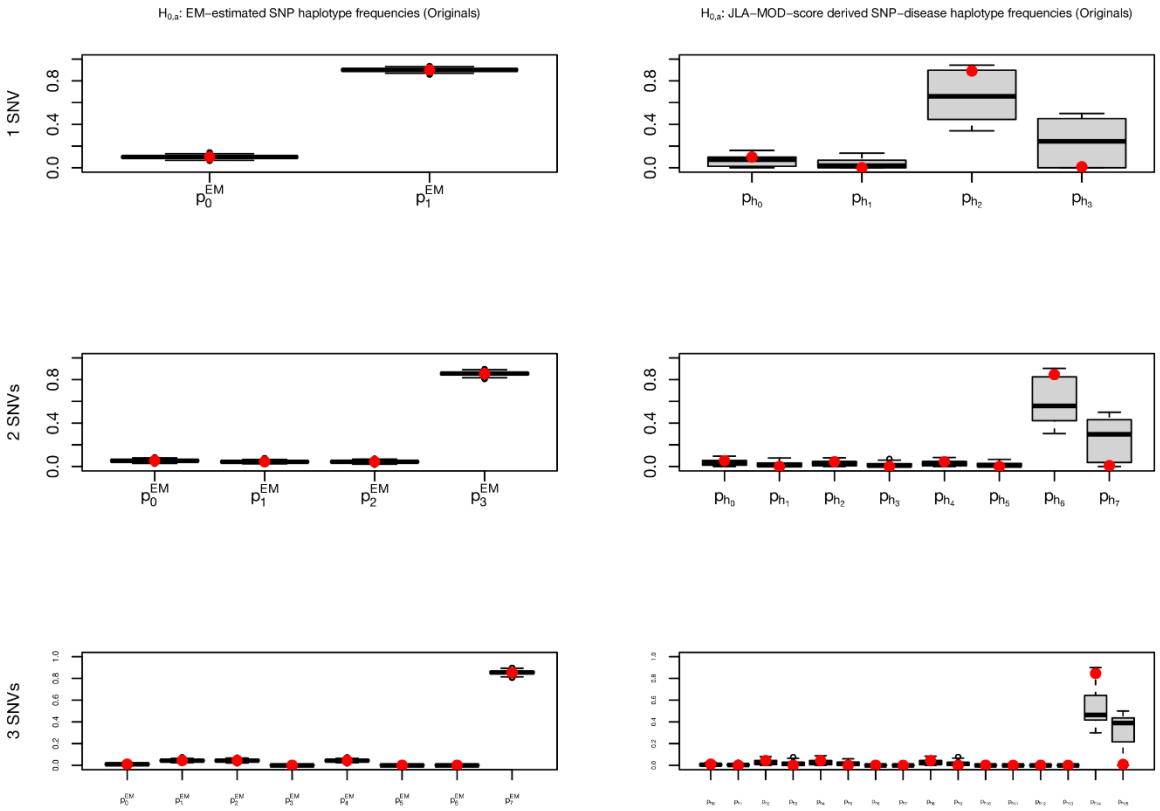
The individual-level data of the FaPaCa study are not publicly available because the data contain sensitive patient data, which underlie data protection rules. This is in accordance with the local ethic vote and the regulations of the FaPaCa registry. Patients' characteristics are available upon request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

References

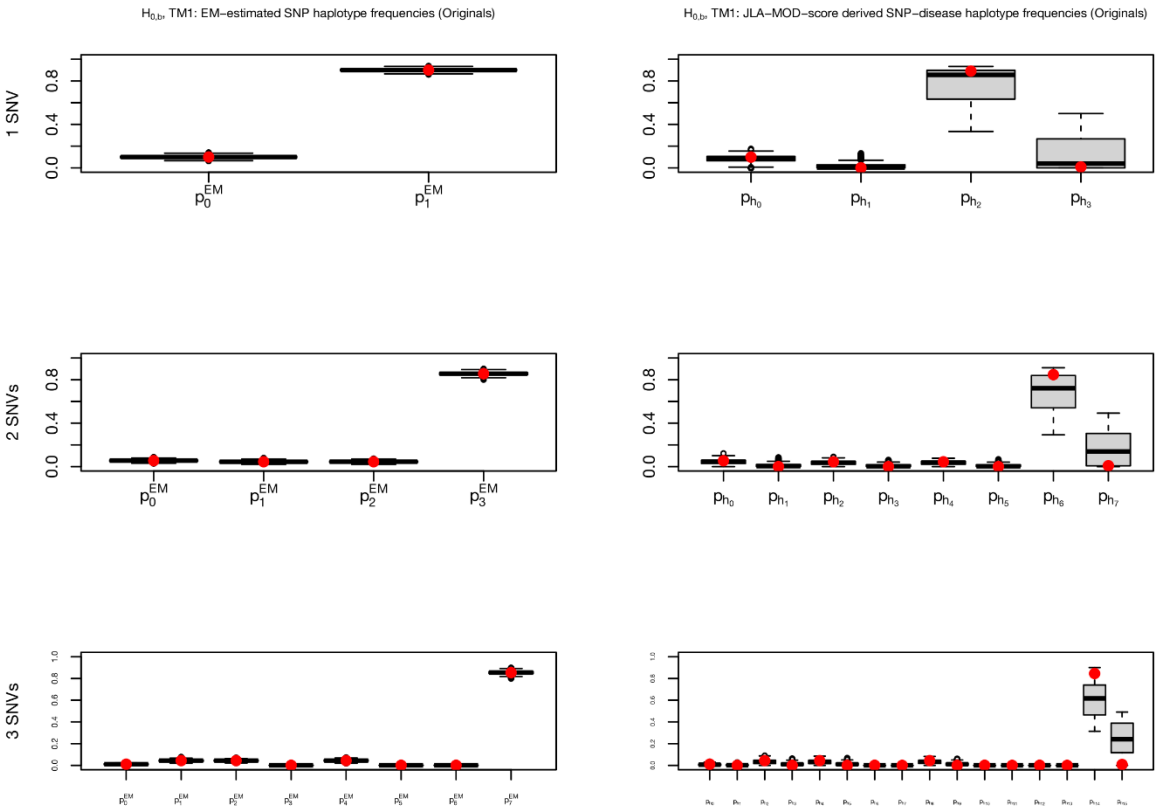
1 Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33(Suppl l):228–37.
2 Terwilliger JD. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one 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, Rosplesczcz 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 1):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.

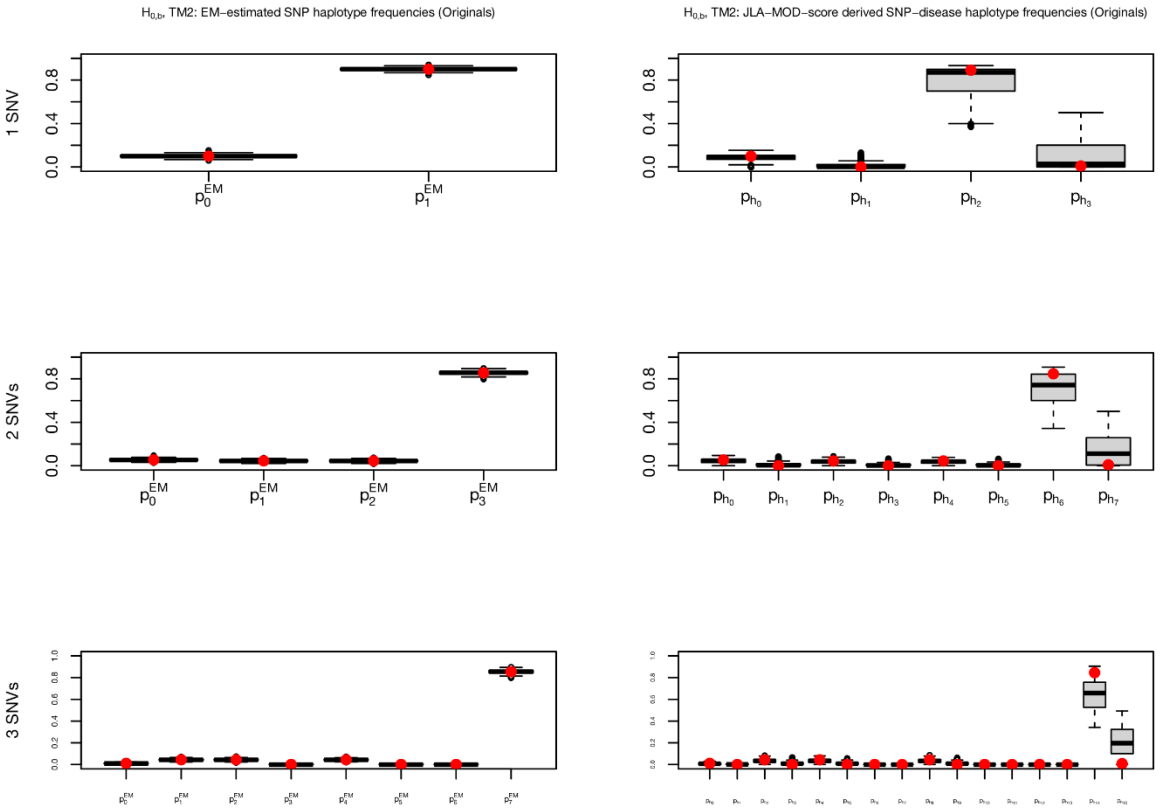
86 Neugebauer M. Mathematische Methoden und Algorithmen zur Analyse genetischer Polymorphismen in Stammbäumen. Inaugural dissertation. University of Bonn; 1989.
87 Powell MJD. A direct search optimization method that models the objective and constraint functions by linear interpolation. In: Gomez S, 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.



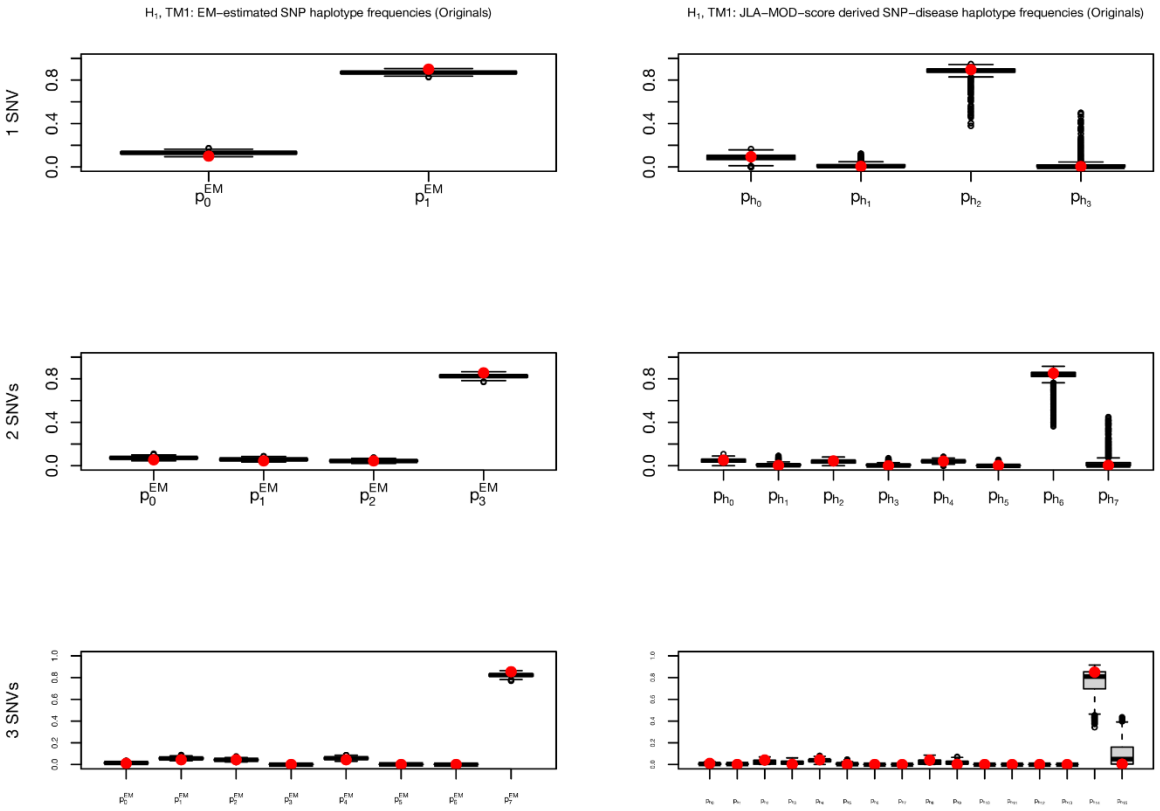
Supplementary Fig. 1: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of no linkage, no association (depiction only for trait model TM1) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. The left column depicts the frequency estimates of the marker haplotypes leaving the disease locus out of the haplotype formulation, which are obtained using the EM algorithm. The right column depicts the frequency estimates of the marker-trait locus haplotypes, which are obtained by maximizing the JLA MOD score likelihood ratio using the COBYLA algorithm. Expected values (red bullets) can be derived from Tables 2 and 3.



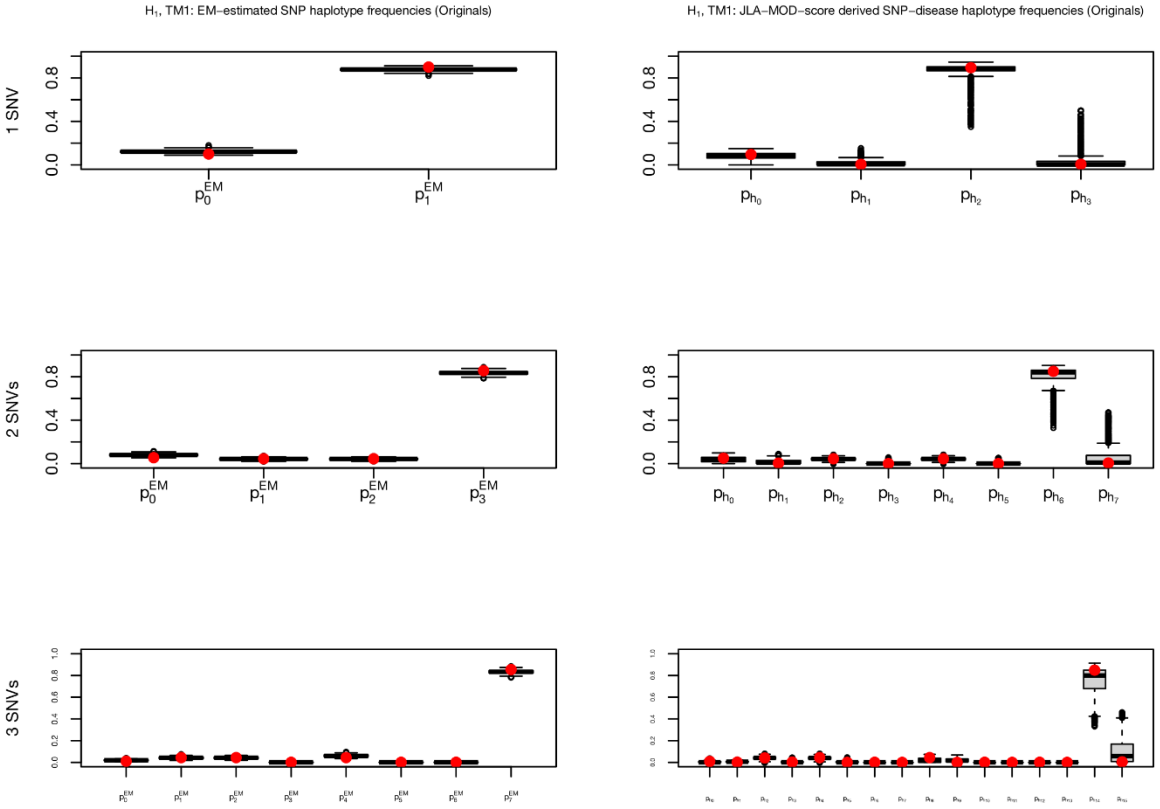
Supplementary Fig. 2: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of linkage, no association (trait model TM1) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. For more details see Supplementary Figure 1.



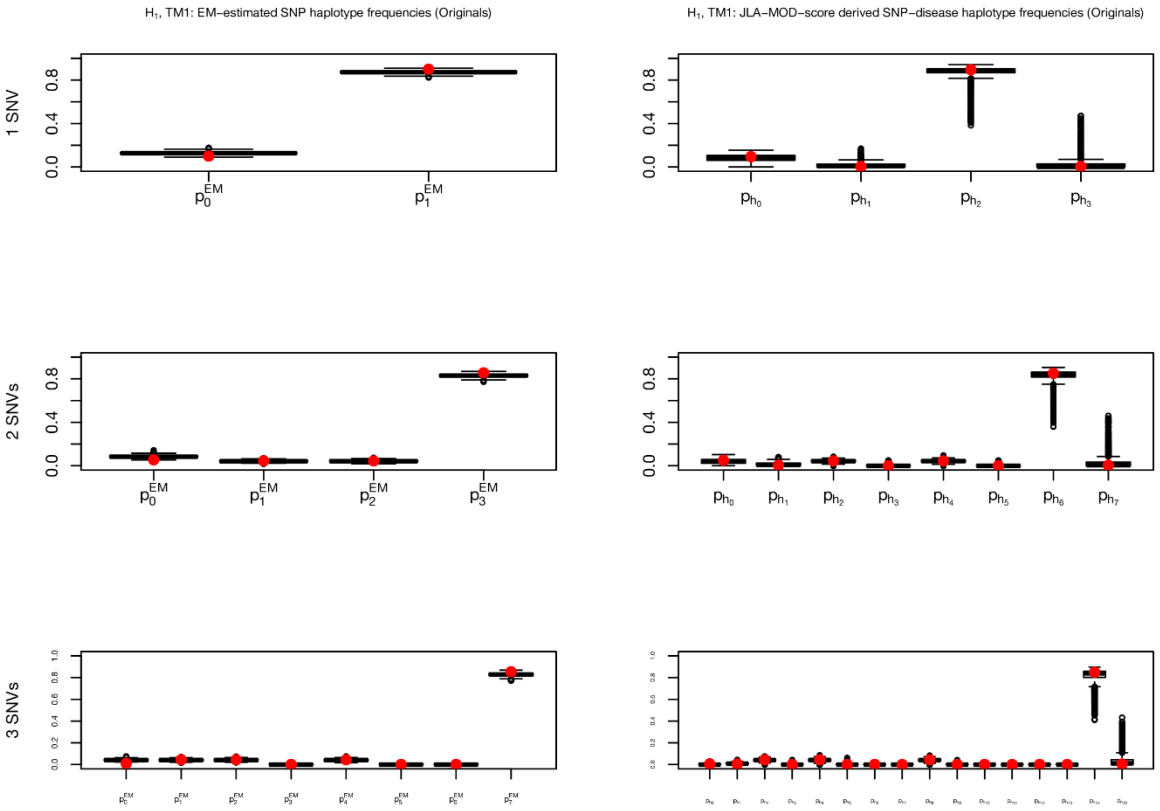
Supplementary Fig. 3: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of linkage, no association (trait model TM2) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. For more details see Supplementary Figure 1.



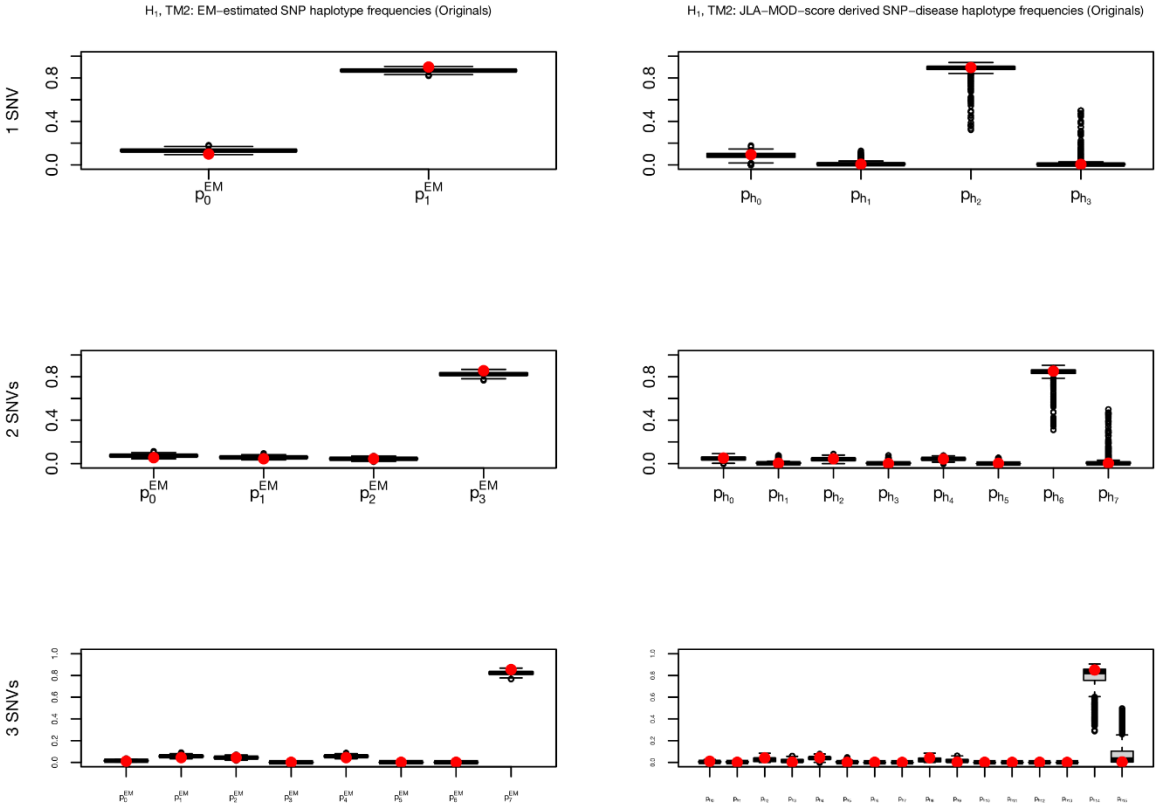
Supplementary Fig. 4: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of linkage and association (trait model TM1, LD pattern S1) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. For more details see Supplementary Figure 1.



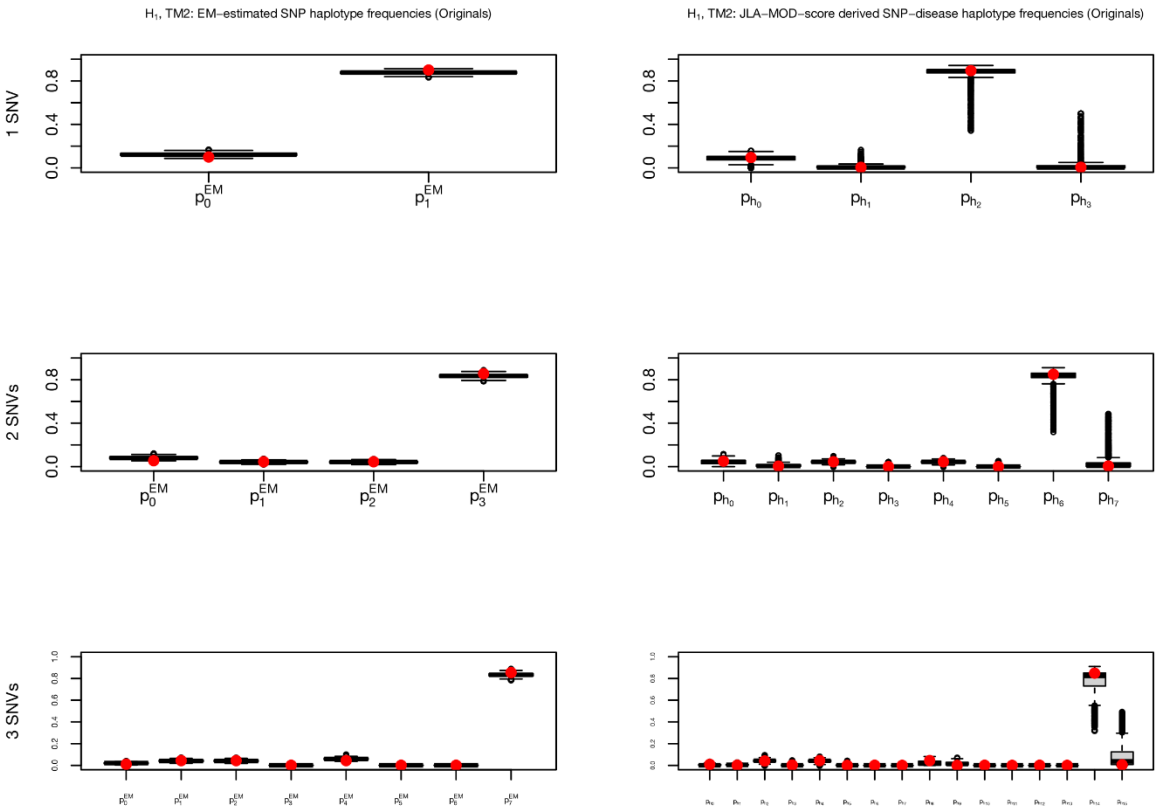
Supplementary Fig. 5: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of linkage and association (trait model TM1, LD pattern S2) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. For more details see Supplementary Figure 1.



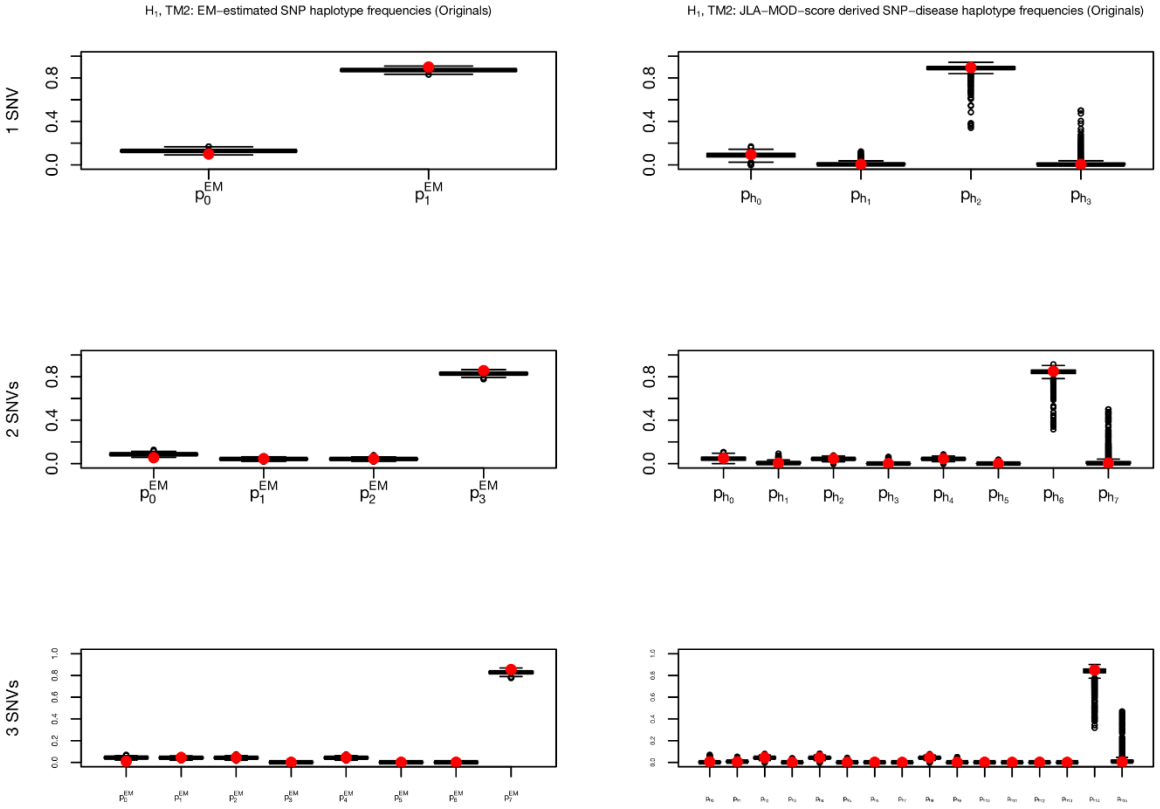
Supplementary Fig. 6: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of linkage and association (trait model TM1, LD pattern S3) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. For more details see Supplementary Figure 1.



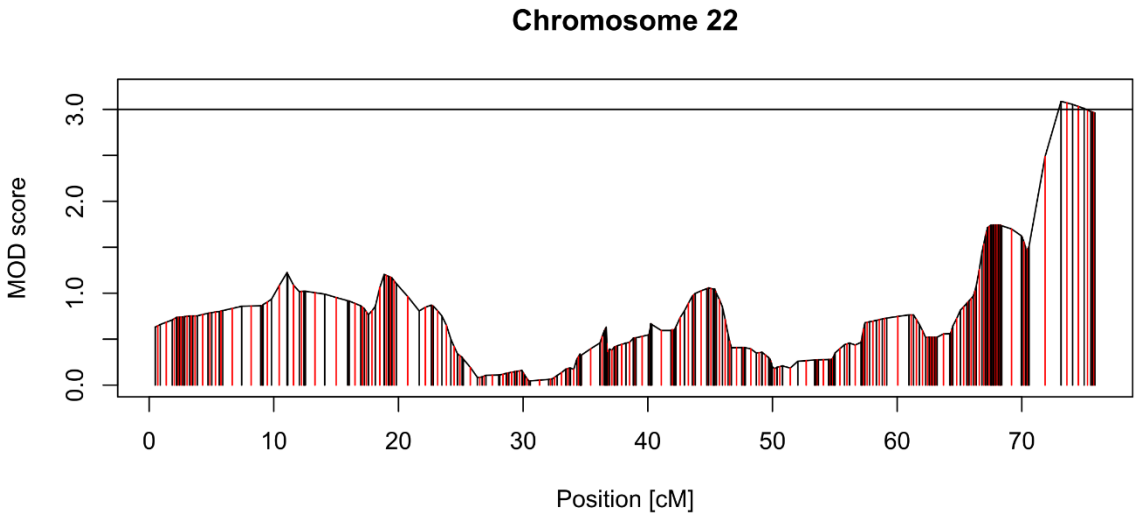
Supplementary Fig. 7: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of linkage and association (trait model TM2, LD pattern S1) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. For more details see Supplementary Figure 1.



Supplementary Fig. 8: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of linkage and association (trait model TM2, LD pattern S2) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. For more details see Supplementary Figure 1.



Supplementary Fig. 9: Boxplots of the empiric distribution of haplotype frequency estimates for the 'original' SLINK datasets simulated under the hypothesis of linkage and association (trait model TM2, LD pattern S3) using either one (row 1), two (row 2), or three (row 3) test SNVs for the GHM JLA analysis. For more details see Supplementary Figure 1.



Supplementary Fig. 10: MOD score curve of chromosome 22. Black vertical lines indicate marker positions, red vertical lines indicate positions halfway between markers. Best-fitting trait models and hence MOD scores were calculated for each investigated position separately (GHM option 'modcalc single').

Supplementary Table 1

TM1		$H_{0,a}$: No linkage, no association							
		Simulated values							
JLA analysis option		$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.09$	$f_{1,mat} = 0.09$	$f_2 = 0.17$	Cramér's $V = 0$	Dominance index $D = 0$	Imprinting index $I = 0$
		Mean (SD)							
1 SNV test marker	Originals	0.27 (0.21)	0.29 (0.32)	0.50 (0.40)	0.50 (0.40)	0.73 (0.39)	0.32 (0.32) [100]	-0.04 (0.63) [99]	0.01 (0.60) [99]
	Replicates	0.26 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.37)	0.32 (0.32) [100]	-0.07 (0.62) [99]	-0.01 (0.62) [99]
2 SNV test markers	Originals	0.30 (0.19)	0.21 (0.29)	0.38 (0.38)	0.38 (0.38)	0.58 (0.42)	0.38 (0.24) [100]	0.0 (0.59) [100]	0.0 (0.63) [100]
	Replicates	0.30 (0.19)	0.22 (0.29)	0.39 (0.38)	0.40 (0.38)	0.60 (0.42)	0.37 (0.24) [100]	-0.01 (0.58) [99]	0.0 (0.64) [99]
3 SNV test markers	Originals	0.37 (0.16)	0.19 (0.25)	0.40 (0.35)	0.39 (0.35)	0.60 (0.39)	0.35 (0.16) [10]	0.03 (0.55) [99]	0.02 (0.61) [99]
	Replicates	0.36 (0.17)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.36 (0.18) [5]	0.0 (0.55) [100]	0.0 (0.63) [100]

TM1		$H_{0,b}$: Linkage, no association							
		Simulated values							
JLA analysis option		$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.09$	$f_{1,mat} = 0.09$	$f_2 = 0.17$	Cramér's $V = 0$	Dominance index $D = 0$	Imprinting index $I = 0$
1 SNV test marker	Originals	0.16 (0.19)	0.09 (0.19)	0.42 (0.38)	0.41 (0.38)	0.74 (0.35)	0.12 (0.21) [100]	-0.02 (0.66) [100]	0.01 (0.46) [100]
	Replicates	0.26 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.37)	0.32 (0.32) [100]	-0.06 (0.62) [99]	-0.01 (0.61) [99]
2 SNV test markers	Originals	0.20 (0.18)	0.10 (0.20)	0.39 (0.37)	0.37 (0.36)	0.64 (0.39)	0.21 (0.19) [100]	0.06 (0.63) [99]	0.01 (0.49) [99]
	Replicates	0.30 (0.19)	0.22 (0.29)	0.39 (0.38)	0.40 (0.38)	0.60 (0.42)	0.37 (0.24) [100]	-0.01 (0.58) [99]	0.0 (0.64) [99]
3 SNV test markers	Originals	0.28 (0.17)	0.07 (0.14)	0.31 (0.32)	0.32 (0.32)	0.57 (0.38)	0.26 (0.09) [9]	0.01 (0.60) [100]	0.0 (0.50) [100]
	Replicates	0.36 (0.17)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.36 (0.18) [5]	0.0 (0.55) [100]	0.0 (0.63) [100]

TM2		$H_{0,b}$: Linkage, no association							
		Simulated values							
JLA analysis option		$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.14$	$f_{1,mat} = 0.04$	$f_2 = 0.17$	Cramér's $V = 0$	Dominance index $D = 0$	Imprinting index $I = 0.625$
1 SNV test marker	Originals	0.13 (0.16)	0.07 (0.18)	0.49 (0.39)	0.25 (0.33)	0.71 (0.37)	0.1 (0.18) [100]	-0.06 (0.57) [100]	0.39 (0.46) [100]
	Replicates	0.26 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.34)	0.32 (0.32) [100]	-0.06 (0.62) [99]	-0.01 (0.61) [99]
2 SNV test markers	Originals	0.18 (0.17)	0.08 (0.19)	0.44 (0.37)	0.24 (0.31)	0.62 (0.39)	0.17 (0.17) [100]	-0.01 (0.55) [99]	0.36 (0.49) [99]
	Replicates	0.30 (0.19)	0.22 (0.29)	0.39 (0.38)	0.40 (0.38)	0.60 (0.42)	0.37 (0.24) [100]	-0.01 (0.58) [99]	0.0 (0.64) [99]
3 SNV test markers	Originals	0.25 (0.16)	0.06 (0.15)	0.36 (0.32)	0.20 (0.27)	0.55 (0.37)	0.26 (0.15) [9]	-0.05 (0.55) [100]	0.33 (0.45) [100]
	Replicates	0.36 (0.17)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.36 (0.18) [5]	0.0 (0.55) [100]	0.0 (0.63) [100]

Supplementary Table 1: Overview of trait-model parameter estimates obtained from JLA MOD score analyses. Values are reported for the null hypothesis of no linkage and no association ($H_{0,a}$, top) for trait model TM1 and linkage, no association ($H_{0,b}$, middle and bottom) for both trait models TM1 and TM2. Values in square brackets indicate the percentage (rounded to the nearest integer) of defined values for Cramér's V and the ratios D and I.
Values for 'Originals' are based on 1000 simulated datasets (SLINK or SLINK-Imprinting), whereas values for 'Replicates' are based on a total of 999000 simulated datasets (GHM-JLA).

Supplementary Table 2

TM1		H_1: Linkage, association							
	SNV scenario	Simulated values							
JLA analysis option	S1	$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.09$	$f_{1,mat} = 0.09$	$f_2 = 0.17$	Cramér's V: 1 SNV: 0.158 2 SNVs: 0.158 3 SNVs: 0.160	Dominance index $D = 0$	Imprinting index $I = 0$
		Mean (SD)							
1 SNV test marker	Originals	0.05 (0.09)	0.05 (0.09)	0.39 (0.37)	0.39 (0.37)	0.70 (0.37)	0.21 (0.20) [100]	0.1 (0.69) [100]	0.01 (0.30) [100]
	Replicates	0.27 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.37)	0.33 (0.33) [100]	-0.08 (0.63) [99]	-0.01 (0.60) [99]
2 SNV test markers	Originals	0.06 (0.1)	0.06 (0.1)	0.39 (0.35)	0.39 (0.35)	0.7 (0.36)	0.25 (0.20) [100]	0.08 (0.66) [100]	0.0 (0.32) [100]
	Replicates	0.3 (0.19)	0.23 (0.29)	0.40 (0.38)	0.40 (0.38)	0.61 (0.42)	0.38 (0.24) [100]	-0.02 (0.59) [99]	0.0 (0.63) [99]
3 SNV test markers	Originals	0.15 (0.14)	0.07 (0.1)	0.34 (0.30)	0.32 (0.28)	0.60 (0.36)	0.47 (0.21) [14]	0.07 (0.61) [100]	0.02 (0.33) [100]
	Replicates	0.36 (0.16)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.37 (0.17) [7]	-0.01 (0.56) [100]	0.0 (0.62) [100]
JLA analysis option	S2	$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.09$	$f_{1,mat} = 0.09$	$f_2 = 0.17$	Cramér's V: 1 SNV: 0.118 2 SNVs: 0.175 3 SNVs: 0.187	Dominance index $D = 0$	Imprinting index $I = 0$
1 SNV test marker	Originals	0.07 (0.12)	0.06 (0.13)	0.41 (0.38)	0.40 (0.37)	0.74 (0.36)	0.20 (0.22) [100]	0.06 (0.69) [100]	0.02 (0.35) [100]
	Replicates	0.26 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.37)	0.33 (0.33) [100]	-0.08 (0.64) [99]	-0.01 (0.6) [99]
2 SNV test markers	Originals	0.09 (0.13)	0.05 (0.11)	0.38 (0.35)	0.36 (0.34)	0.68 (0.38)	0.27 (0.21) [100]	0.07 (0.66) [100]	0.01 (0.33) [100]
	Replicates	0.30 (0.19)	0.22 (0.29)	0.40 (0.38)	0.40 (0.38)	0.61 (0.42)	0.38 (0.24) [100]	-0.02 (0.59) [99]	0.0 (0.64) [99]
3 SNV test markers	Originals	0.14 (0.14)	0.05 (0.08)	0.32 (0.29)	0.32 (0.28)	0.60 (0.35)	0.42 (0.19) [9]	0.04 (0.62) [100]	0.01 (0.32) [100]
	Replicates	0.36 (0.16)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.37 (0.17) [5]	0.0 (0.56) [100]	0.0 (0.63) [100]
JLA analysis option	S3	$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.09$	$f_{1,mat} = 0.09$	$f_2 = 0.17$	Cramér's V: 1 SNV: 0.141 2 SNVs: 0.201 3 SNVs: 0.474	Dominance index $D = 0$	Imprinting index $I = 0$
1 SNV test marker	Originals	0.06 (0.11)	0.05 (0.10)	0.38 (0.36)	0.38 (0.36)	0.71 (0.37)	0.21 (0.22) [100]	0.05 (0.68) [100]	0.0 (0.32) [100]
	Replicates	0.26 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.37)	0.33 (0.33) [100]	-0.08 (0.83) [99]	-0.01 (0.60) [99]
2 SNV test markers	Originals	0.06 (0.11)	0.05 (0.09)	0.36 (0.34)	0.35 (0.33)	0.67 (0.37)	0.28 (0.22) [100]	0.07 (0.66) [100]	0.01 (0.29) [100]
	Replicates	0.30 (0.19)	0.23 (0.29)	0.40 (0.38)	0.40 (0.38)	0.61 (0.42)	0.38 (0.24) [100]	-0.02 (0.59) [99]	0.0 (0.64) [99]
3 SNV test markers	Originals	0.06 (0.09)	0.04 (0.07)	0.33 (0.28)	0.32 (0.28)	0.61 (0.35)	0.54 (0.19) [11]	0.13 (0.61) [100]	0.01 (0.24) [100]
	Replicates	0.36 (0.16)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.38 (0.17) [6]	0.0 (0.56) [100]	0.0 (0.63) [100]

Supplementary Table 2: Overview of trait-model parameter estimates obtained from JLA MOD score analyses. Values are reported for the null hypothesis of linkage and association (H_1) and trait model TM1. For more information see Supplementary Table 1.

Supplementary Table 3

TM2		H_1: Linkage, association							
	SNV scenario	Simulated values							
JLA analysis option	S1	$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.14$	$f_{1,mat} = 0.04$	$f_2 = 0.17$	Cramér's V: 1 SNV: 0.158 2 SNVs: 0.158 3 SNVs: 0.160	Dominance index $D = 0$	Imprinting index $I = 0.625$
		Mean (SD)							
1 SNV test marker	Originals	0.03 (0.08)	0.04 (0.06)	0.49 (0.39)	0.14 (0.18)	0.69 (0.38)	0.21 (0.18) [100]	-0.10 (0.47) [100]	0.55 (0.35) [100]
	Replicates	0.27 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.37)	0.33 (0.33) [100]	-0.08 (0.64) [99]	-0.01 (0.61) [99]
2 SNV test markers	Originals	0.04 (0.08)	0.05 (0.08)	0.52 (0.38)	0.16 (0.18)	0.70 (0.36)	0.24 (0.18) [100]	-0.09 (0.45) [100]	0.54 (0.38) [100]
	Replicates	0.30 (0.19)	0.23 (0.29)	0.40 (0.38)	0.40 (0.38)	0.61 (0.42)	0.38 (0.24) [100]	-0.03 (0.59) [99]	0.0 (0.63) [99]
3 SNV test markers	Originals	0.12 (0.13)	0.06 (0.08)	0.42 (0.31)	0.16 (0.17)	0.58 (0.35)	0.47 (0.21) [15]	-0.05 (0.44) [100]	0.5 (0.37) [100]
	Replicates	0.36 (0.16)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.37 (0.17) [7]	-0.01 (0.56) [100]	0.0 (0.62) [100]
JLA analysis option	S2	$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.14$	$f_{1,mat} = 0.04$	$f_2 = 0.17$	Cramér's V: 1 SNV: 0.118 2 SNVs: 0.175 3 SNVs: 0.187	Dominance index $D = 0$	Imprinting index $I = 0.625$
1 SNV test marker	Originals	0.05 (0.10)	0.04 (0.09)	0.52 (0.40)	0.16 (0.21)	0.74 (0.35)	0.15 (0.16) [100]	-0.13 (0.51) [100]	0.51 (0.40) [100]
	Replicates	0.26 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.37)	0.33 (0.33) [100]	-0.08 (0.63) [99]	-0.01 (0.61) [99]
2 SNV test markers	Originals	0.06 (0.10)	0.05 (0.09)	0.47 (0.37)	0.15 (0.19)	0.67 (0.37)	0.24 (0.18) [100]	-0.12 (0.46) [100]	0.52 (0.4) [100]
	Replicates	0.30 (0.19)	0.22 (0.29)	0.40 (0.38)	0.40 (0.38)	0.61 (0.42)	0.38 (0.24) [100]	-0.02 (0.59) [99]	0.0 (0.64) [99]
3 SNV test markers	Originals	0.11 (0.13)	0.05 (0.07)	0.42 (0.33)	0.16 (0.19)	0.60 (0.35)	0.42 (0.16) [10]	-0.07 (0.45) [100]	0.49 (0.39) [100]
	Replicates	0.36 (0.16)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.37 (0.17) [5]	0.0 (0.56) [100]	0.0 (0.63) [100]
JLA analysis option	S3	$p_m = 0.01$	$f_0 = 0.01$	$f_{1,pat} = 0.14$	$f_{1,mat} = 0.04$	$f_2 = 0.17$	Cramér's V: 1 SNV: 0.141 2 SNVs: 0.201 3 SNVs: 0.474	Dominance index $D = 0$	Imprinting index $I = 0.625$
1 SNV test marker	Originals	0.03 (0.07)	0.04 (0.08)	0.53 (0.40)	0.16 (0.21)	0.72 (0.37)	0.17 (0.16) [100]	-0.09 (0.49) [100]	0.55 (0.35) [100]
	Replicates	0.26 (0.21)	0.28 (0.32)	0.48 (0.39)	0.49 (0.40)	0.74 (0.37)	0.33 (0.33) [100]	-0.08 (0.63) [99]	-0.01 (0.60) [99]
2 SNV test markers	Originals	0.04 (0.08)	0.04 (0.09)	0.49 (0.37)	0.15 (0.20)	0.69 (0.37)	0.25 (0.18) [100]	-0.11 (0.45) [100]	0.52 (0.38) [100]
	Replicates	0.30 (0.19)	0.23 (0.29)	0.40 (0.38)	0.40 (0.38)	0.61 (0.42)	0.38 (0.24) [100]	-0.02 (0.60) [99]	0.0 (0.64) [99]
3 SNV test markers	Originals	0.05 (0.08)	0.04 (0.06)	0.48 (0.33)	0.15 (0.17)	0.64 (0.34)	0.57 (0.20) [12]	-0.04 (0.41) [100]	0.55 (0.33) [100]
	Replicates	0.36 (0.16)	0.19 (0.25)	0.38 (0.35)	0.38 (0.35)	0.59 (0.39)	0.38 (0.17) [6]	0.0 (0.56) [100]	0.0 (0.63) [100]

Supplementary Table 3: Overview of trait-model parameter estimates obtained from JLA MOD score analyses. Values are reported for the null hypothesis of linkage and association (H_1) and trait model TM2. Values in square brackets indicate the percentage (rounded to the nearest integer) of defined values for Cramér's V and the ratios D and I.

7. Literaturverzeichnis

Abecasis GR, Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet. 2005;77(5):754—67.

Bartsch DK, Matthäi E, Mintziras I, Bauer C, Figiel J, Sina-Boemers M, Gress TM, Langer P, Slater EP. The German National Case Collection for Familial Pancreatic Cancer (FaPaCa)—knowledge gained in 20 years. Dtsch Arztebl Int. 2021;118(10):163—8.

Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 Suppl:228—37.

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):20180025.

Brugger M, Lutz M, Müller-Nurasyid M, Lichtner P, Slater EP, Matthäi E, Bartsch DK, Strauch K. Joint linkage and association analysis with GENEHUNTER-MODSCORE with an application to familial pancreatic cancer. Hum Hered. 2024;89(1):8—31.

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.

Brugger M, Strauch K. Fast linkage analysis with MOD scores using algebraic calculation. Hum Hered. 2014;78(3-4):179—94.

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(4):363—372.

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.

Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J. Effects of misspecifying genetic parameters in lod score analysis. Biometrics. 1986;42(2):393—9.

Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the E-M algorithm. J R Stat Soc Series B. 1977;39(1):1—38.

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.

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.

Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Hum Hered. 1971;21(6):523—42.

Fan R, Jung J. High-resolution joint linkage disequilibrium and linkage mapping of quantitative trait loci based on sibship data. Hum Hered. 2003;56(4):166—87.

Gertz EM, Hiekkalinna T, Digabel SL, Audet C, Terwilliger JD, Schäffer AA. PSEUDOMARKER 2.0: efficient computation of likelihoods using NOMAD. BMC Bioinformatics. 2014;15:47.

Ginsburg E, Malkin I, Elston RC. Sampling correction in linkage analysis. Genet Epidemiol. 2004;27(2):87—96.

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.

Hall JG. Genomic imprinting: review and relevance to human diseases. Am J Hum Genet. 1990;46(5):857—73.

Handel-Fernandez ME, Nassiri M, Arana M, Perez MM, Fresno M, Nadji M, Vincek V. Mapping of genetic deletions on the long arm of chromosome 22 in human pancreatic adenocarcinomas. Anticancer Res. 2000;20(6B):4451—6.

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.

Johnson SG. The NLopt nonlinear-optimization package. 2020. http://github.com/stevengj/nlopt

Jorde LB. Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet. 1995;56(1):11—4.

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.

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.

Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A. 1987;84(8):2363—7.

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.

Malkin I, Elston RC. Response to letter by Veronica J. Vieland and Susan E. Hodge. Genet Epidemiol. 2005;28(3):286—7.

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.

Morton NE. Sequential tests for the detection of linkage. Am J Hum Genet. 1955;7(3):277—318.

Ott J. Analysis of human genetic linkage (third edition). Johns Hopkins University Press, Baltimore. 1999.

Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet. 2015;16(5):275—84.

Portin P, Wilkins A. The evolving definition of the term "gene". Genetics. 2017;205(4):1353—1364.

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. S. 51—67.

Powell MJD. Direct search algorithms for optimization calculations. Acta Numer. 1998;7:287—336.

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.

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.

Ruddle FH. The William Allan Memorial Award address: Reverse genetics and beyond. Am J Hum Genet. 1984;36(5):944—953.

Smalley SL. Sex-specific recombination frequencies: a consequence of imprinting? Am J Hum Genet. 1993;52(1):210—2.

Strauch K. Parametric linkage analysis with automatic optimization of the disease model parameters. Am J Hum Genet. 2003;73(suppl 1):A2624.

Strauch K. MOD-score analysis with simple pedigrees: an overview of likelihood-based linkage methods. Hum Hered. 2007;64(3):192—202.

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.

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.

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.

Weiss K M. Genetic variation and human disease: principles and evolutionary approaches. Cambridge University Press, Cambridge. 1993.

Anhang A: Paper III

Brugger M, Rospleszcz S, Strauch K. Estimation of trait-model parameters in a MOD score linkage analysis. Hum Hered. 2016;82(3-4):103—139.

Original Paper

Human Heredity

Hum Hered 2016;82:103–139
DOI: 10.1159/000479738

Received: March 27, 2017
Accepted: July 25, 2017
Published online: November 1, 2017

Estimation of Trait-Model Parameters in a MOD Score Linkage Analysis

Markus Brugger Susanne Rospleszcz Konstantin Strauch

Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, and Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Germany

Keywords
Parametric linkage analysis · MOD scores · Trait-model parameters · Identifiability · Estimation bias

Abstract
Background/Aims: Theoretically, the trait-model parameters (disease allele frequency and penetrance function) can be estimated without bias in a MOD score linkage analysis. We aimed to practically evaluate the MOD score approach regarding its ability to provide unbiased trait-model parameters for various pedigree-type and trait-model scenarios. We further investigated the ability of the MOD score approach to detect imprinting using affected sib pairs (ASPs) and affected half-sib pairs (AHSPs) when all parental genotypes are missing. ***Methods:*** Simulated pedigree data were analyzed using the GENEHUNTER-MODSCORE software package. Parameter estimation performance in terms of bias and variability was evaluated with regard to trait-model type and pedigree complexity. ***Results:*** Generally, parameters were estimated with lower bias and variability with increasing pedigree complexity, especially for recessive and overdominant models. However, dominant and additive models could hardly be distinguished even when using 3-generation pedigrees. Imprinting could clearly be detected for mixtures of mainly ASPs and only few AHSPs with the common parent of the imprinted sex, even though no parental genotypes were available. ***Conclusion:*** Our results provide guidance to researchers regarding the possibility to estimate trait-model parameters by a MOD score analysis, including the degree of imprinting, with certain types of pedigrees.

Introduction

Trait Inheritance and Pedigree Analysis

The inheritance of a trait is defined as the mechanism by which the joint phenotypic distribution of the particular trait in pedigree members can explicitly be described [1]. A pedigree can be considered as a discrete unit of a population for which the relationship connecting any pair of pedigree members is unambiguously known. There is hence no other individual for which a relationship to any of these pedigree members can be established. Under the assumption that pedigrees implicitly contain information about details of the mode of inheritance of a trait through the covariation and cosegregation of the

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

Markus Brugger
Institute of Genetic Epidemiology
Helmholtz Zentrum München – German Research Center for Environmental Health
Ingolstädter Landstrasse 1, DE–85764 Neuherberg (Germany)
E-Mail markus.brugger@helmholtz-muenchen.de

trait characteristics among its members, collecting and analyzing samples of pedigrees can be used to study the trait inheritance. In genetics, inference about trait inheritance by pedigree analysis is made assuming that the main factors underlying the inheritance are genes. Mathematical-genetic models can then be used to describe the trait inheritance, and these models are tested using pedigree samples drawn from the population.

If the genetic model of trait inheritance is inferred on the basis of the pedigree sample, which contains the necessary information through the joint phenotypic cosegregation in the pedigree members, such an analysis is called "segregation analysis" [1]. If the purpose of the analysis is to map the putative disease gene(s), whose existence may have been previously established by segregation analysis to specific chromosomal segments by investigation of cosegregation of DNA marker alleles and the trait phenotype, such an analysis is called "linkage analysis" [1]. Nowadays, pure segregation analysis is of less practical importance than it has been a few decades ago. With increasing availability of DNA marker maps and rapid and cost-effective DNA genotyping techniques, linkage analysis has become the state-of-the-art technique of pedigree analysis. In addition, association analysis can be performed with pedigrees as well as samples of unrelated individuals. However, software packages for segregation analysis like PAP [2], S.A.G.E. [3], and MORGAN [4, 5] continue to be available and provide great flexibility with respect to fitting the model for the mode of inheritance (see also e.g. Kriszt et al. [6] for a recent publication using complex segregation analysis with keratoconus pedigrees).

Linkage Analysis

In earlier times, linkage analysis has been used to map genes that were already known to exist. In the meantime, linkage analysis serves 2 purposes: (1) to prove the existence of a disease gene and (2) to map it [7]. Linkage analysis methods can be distinguished as model-based or model-free [8]. The former is also known as parametric or LOD score linkage analysis for which a certain set of trait-model parameters regarding the segregation of the disease is explicitly assumed in the genetic likelihood. The latter, which is also known as nonparametric linkage analysis, proceeds without such explicit models. These 2 types of linkage analysis are, however, closely related to each other. It can be shown that certain nonparametric and parametric linkage tests are equivalent for any type of pedigree [9, 10] and can be considered as different ways to parametrize the allele-sharing probabilities, i.e., the probabilities of allele(s) shared identical-by-descent (IBD) by affected pedigree members, in the genetic likelihood.

Mode of Inheritance and Trait-Model Parameters

A crucial factor in linkage analysis is the true mode of inheritance. Under the term "mode of inheritance," 2 concepts are often subsumed that need, however, to be distinguished. The first concept is the genetic mechanism of the disease involving the number of loci, the number of alleles at each locus, and the segregation parameters including the recombination fraction among the trait loci as well as between them and any marker(s) [11]. The second concept is the genotype-phenotype relation, which is defined by the penetrance function, i.e., the probability that an individual with a certain number of copies of the disease allele is affected by the disease. The genetic mechanism of the disease, apart from the recombination fraction, is assumed to be known for linkage analysis. In the case of a binary trait governed by a single diallelic autosomal locus, which is assumed throughout this paper, the disease allele frequency p and the 3 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, can be subsumed under the term "trait-model parameters." In the case of parametric linkage analysis, 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 (JSL) analysis. A specific type of this approach is the MOD score analysis, which was first proposed by Risch [12]. If the genetic mechanism of the trait is not modelled correctly, however, which is expected in practice due to the large number of possible inheritance modes, parameter estimates obtained from a MOD score analysis will be asymptotically biased [11, 13].

Likelihood and Sample Space

In pedigree analysis, the likelihood given a particular sample of pedigrees can be defined as the probability to observe the data available for the individuals in the pedigree, constructed under a certain genetic model. In fact, any formulation that is proportional to this probability can be used as the likelihood. The pedigree samples used for pedigree analysis are collected from what is called the "real" population that is defined on the basis of usually unknown factors like the population's origin and history. This real population is mapped into a set of disjoint pedigrees by the use of those relationships between members of the real population that can unambiguously be established [1]. These disjoint pedigrees are then further deter-

mined by the predefined sampling design, which partitions the pedigrees into substructures of certain inheritance relations, e.g., sibships with all other relationships outside sibships being ignored. The resulting structures are called "true pedigrees." As described in Ginsburg et al. [1], pedigree analysis is performed on sampled pedigrees collected from the set of true pedigrees. The subset of pedigrees that in principle can be sampled according to the sampling design is called the "sample space." The sampling procedure involves the pedigree ascertainment (primary selection), the intrafamilial extension (inclusion of additional relatives), and the selective inclusion in the analysis (censoring).

In the following, we will assume that ascertainment takes place through probands. For each true pedigree, there are members who could "potentially" become probands due to prespecified proband characteristics, e.g., geographic area, age, sex, but independently of their phenotypes. This subset of potential probands in the true pedigree, including both their relationships and phenotypes, is called the "proband sampling frame" (PSF, [14]). It can be shown that assuming the wrong mode of inheritance and/or the wrong model for the sampling procedure leads to asymptotically biased trait-model parameters and nuisance parameters of the sampling model when performing maximum likelihood estimation [15]. In order to obtain unbiased parameter estimates, the pedigree likelihood is defined as the probability of the particular pedigree data having been sampled (ascertained, extended, and included in the analysis) on the sample space generated by the sampling procedure under the given mode of inheritance [1]. The sample space for the given sampling procedure is the probability that at least 1 pedigree is sampled from the set of true pedigrees [1]. In this general form, however, the pedigree likelihood cannot be calculated using only the sampled data [1]. This would demand knowledge about the distribution of possible PSFs to calculate the sample space on which the likelihood is defined. Therefore, pedigree likelihoods are conditioned on specific parts of the sampled data to circumvent this problem and – by the same token – to retain unbiasedness of parameter estimates. In the following sections, pedigree likelihoods, which are conditioned on specific parts of the sampled data, are briefly introduced in the context of JSL analysis.

Sampling Model-Based Likelihood

As was explained in the previous section, the pedigree likelihood provides consistent estimates of the trait-model parameters if it is conditioned on the pedigree having been sampled, i.e., ascertained, extended, and included in the sample under analysis [16]. This also holds true for JSL analysis. In parametric JSL analysis, which is the main focus of this paper, the likelihood is formulated using the trait-model parameters, i.e., the disease allele frequency p and the penetrances f_0, f_1, and f_2, as well as the marker allele frequencies and the recombination fractions – and, if applicable, linkage disequilibria (LD) between loci. These parameters can be subsumed under the term "joint trait-marker inheritance parameters" [16]. In addition, information about the following aspects must also be included in the likelihood: (1) the whole PSF structure and its population distribution, which is relevant for ascertainment, (2) the pedigree extension procedure, and (3) the conditions relevant to inclusion, which could be specific marker genotypes of certain pedigree members. Since the population distribution of the PSF structure is unknown, the pedigree likelihood can be conditioned on the substructure of the pedigree that is "relevant to sampling" (RS), in order to make the likelihood calculable and to properly take the sampling procedure into account. The structure RS corresponds to all PSF members of the true pedigree under study – i.e., the part of the pedigree "relevant to ascertainment" (RA) – and those pedigree members responsible for the inclusion of the pedigree in the sample. Importantly, the likelihood is only conditioned on the structure RS but not on the phenotypes of the corresponding pedigree members. Since the likelihood includes explicit details of the sampling procedure, it is termed "sampling model-based (SMB) likelihood" [16]. The SMB likelihood provides asymptotically unbiased estimates of all joint trait-marker inheritance parameters, including the mode of inheritance, as well as of the parameters determining the ascertainment, extension, and inclusion procedure [1].

Sampling Model-Free Likelihood

A sampling model-free (SMF) likelihood can be formulated using a more robust procedure initially proposed by Ewens and Shute [17] in the context of segregation analysis, in which uncertainties about the ascertainment procedure are controlled by conditioning the likelihood on that part of the pedigree data RA. The latter approach is called "ascertainment assumption-free" (AAF) and can readily be extended to be SMF, if the likelihood is also conditioned on that part of the data RS [16]. The part of the data RS is the data RA and that part of the data relevant to inclusion, which could be, e.g., certain parental marker genotypes. In contrast to the SMB likelihood, which is conditioned only on the structure RS, the SMF

likelihood is conditioned on the data RS, i.e., structure as well as marker and trait values RS. This SMF likelihood provides asymptotically unbiased estimates of all joint trait-marker inheritance parameters, including the mode of inheritance, as well as of the extension parameter [1].

Likelihood in a MOD Score Analysis

The question arises which kind of likelihood underlies a JSL analysis using the MOD score, and if it is in principle possible to obtain unbiased parameters from this procedure. As shown by Clerget-Darpoux et al. [18] and later also by Elston [11], maximizing the LOD score in the context of a MOD score analysis is equivalent to maximizing the likelihood of the marker data, conditional on the pedigree structure and conditional on all the trait data, i.e., not only on that part RS. This conditional likelihood – from now on referred to as "MOD score likelihood" – does not depend on the ascertainment scheme, provided that the sampling of pedigrees is independent of marker data. Hence, this means that selective inclusion of pedigrees based on marker genotypes (i.e., marker-dependent sampling) is not controlled in the MOD score likelihood, because it does not contain the inclusion parameter. As a consequence, the MOD score will yield biased estimates of the joint trait-marker inheritance parameters if there is association between disease and marker alleles (LD > 0), because ascertainment is no longer marker-independent in that case [19].

The following conditions must be satisfied to obtain unbiased estimates of the joint trait-marker inheritance parameters from a MOD score analysis [1, 19]: (i) the marker locus must be truly linked to the trait locus, (ii) the genetic mechanism of the trait (number of loci and number of alleles at each locus) is known, (iii) sampling is marker-independent, (iv) the model for the pedigree extension procedure is known, and either (v) trait values are available for all members of the PSF, which has to be completely known, or (vi) the ascertainment is proband-independent (PI) or single in the sense described by Hodge and Vieland [20], i.e., all pedigrees have equal probabilities of being ascertained, independent of pedigree size or structure, or (vii) the joint probability of the unobserved trait phenotypes of the members of the PSF, conditional on the trait and marker phenotypes of all the observed pedigree members, does not depend on the marker phenotypes. Condition (v) reflects that the MOD score likelihood can be derived from the SMB likelihood by conditioning the latter on the trait values of all individuals, including all PSF members, in addition to the structure RS. Condition (vi) is due to the fact that the MOD score likelihood does not include an ascertainment parameter as opposed to the SMB likelihood, which contains such a parameter. The probability of ascertainment, however, actually depends on the joint trait-marker inheritance parameters, if sampling is not PI or single [21]. Only with PI or single ascertainment, the probability of ascertainment no longer depends on these parameters and can, therefore, be omitted in the likelihood without influencing the estimates of the parameters [1]. Without specifying details of the sampling procedure, parameter estimates are also consistent when missing trait values of the PSF members do not depend on marker phenotypes (condition [vii]). However, this only holds in the case of no LD and no linkage between trait and marker locus, or if the trait phenotype unambiguously defines the trait genotype [19].

The MOD score likelihood differs from the SMF likelihood by the fact that it is conditioned on all trait values (i.e., not only of the PSF members) in addition to the data RS, and that it assumes PI or single ascertainment as well as marker-independent sampling, rather than specifying some value for the ascertainment probability in the likelihood. This is why the MOD score likelihood can be considered to be somewhere between SMB and SMF. If sampling is marker-independent, but conditions (i), (ii), and (iv) are not simultaneously satisfied, parameter estimates obtained from MOD score analyses will be biased. If conditions (i)–(iv) hold, but neither condition (v), (vi), nor (vii) is met, the estimate of the recombination frequency will only slightly be biased [1]. In this case, it is of note that estimates of the recombination fraction are biased even when trait-model parameters are fixed at their true values [22].

Summary of Conditions to Obtain Unbiased Parameter Estimates from a MOD Score Analysis

The pedigree likelihood of the MOD score approach delivers asymptotically unbiased estimates of the joint trait-marker inheritance parameters (recombination fraction, allele frequency, and penetrances, but not the LD parameter), if the following conditions are satisfied (see also Malkin and Elston [19]):

i The marker is truly linked.
AND
ii The genetic mechanism of the trait (number of loci and number of alleles at each locus) is known.
AND
iii Sampling (ascertainment, extension, inclusion) of pedigrees is independent from marker data.
AND

iv The model of extension is known.
AND
At least 1 of the following 3 conditions is satisfied:
v All members of the pedigree PSF must have measured trait values (if not sampled, information on trait values can be gathered using a questionnaire as proposed by Ginsburg et al. [16]).
OR
vi The ascertainment procedure is PI or single in the sense of Hodge and Vieland [20].
OR
vii The joint probability of the unobserved trait phenotypes of the members of the PSF, conditional on the trait and marker phenotypes of all the observed pedigree members, does not depend on the marker phenotypes.

Hence, unbiased estimates of the joint trait-marker inheritance parameters can in principle be obtained without explicitly formulating the ascertainment and inclusion procedures. It should further be noted that the likelihood correction in a MOD score analysis directly follows from the AAF method proposed in Ewens and Shute [17]. Whereas conditions (i)–(iii) are crucial, conditions (v)–(vii) may be of minor impact on the bias of parameter estimates in practice [20, 23]. With respect to condition (v), if members of the pedigree PSF are not sampled and trait values cannot be gathered using a questionnaire, an approximate likelihood using the sample mean of the trait value can be constructed [1]. Condition (iv) could be satisfied as follows. PI sampling implies that fixed pedigree structures are sampled, which renders a specification of the extension parameter pointless. With single ascertainment, the pedigree extension model could be chosen to be trait-independent, such that any initially sampled subpedigree is further extended using all available relatives, regardless of their phenotypes and with a random, trait-independent stopping rule. If this holds true, an extension parameter does not have to be formulated in the likelihood. Despite being hard to achieve in practice, conditions (iv)–(vii) can in theory be fulfilled. If not, the resulting bias in parameter estimates is argued to be small [20], but numerical quantification of the bias of the joint trait-marker inheritance parameters obtained from a MOD score analysis under many different sampling schemes is not available so far. This would demand an extensive simulation study to prove that the MOD score approach is robust with regard to its ability to estimate parameters, even if some necessary assumptions do not hold. Even if all necessary conditions are satisfied, a bias of maximum likelihood estimates can nevertheless occur for finite sample sizes. In addition, variances of the obtained estimates are expected to be rather large using the MOD score likelihood due to a loss of pedigree information by conditioning not only on the pedigree structure but also on the trait data of all individuals [24].

The focus of the present paper is the proof-of-principle of the ability of a MOD score analysis to obtain asymptotically unbiased joint trait-marker inheritance parameters in practice, given that conditions (i)–(iv) and at least one of (v)–(vii) are satisfied. In particular, the identifiability (see also next section) of these parameters using various pedigree types and realistic sample sizes will be investigated.

Identifiability of Inheritance Parameters

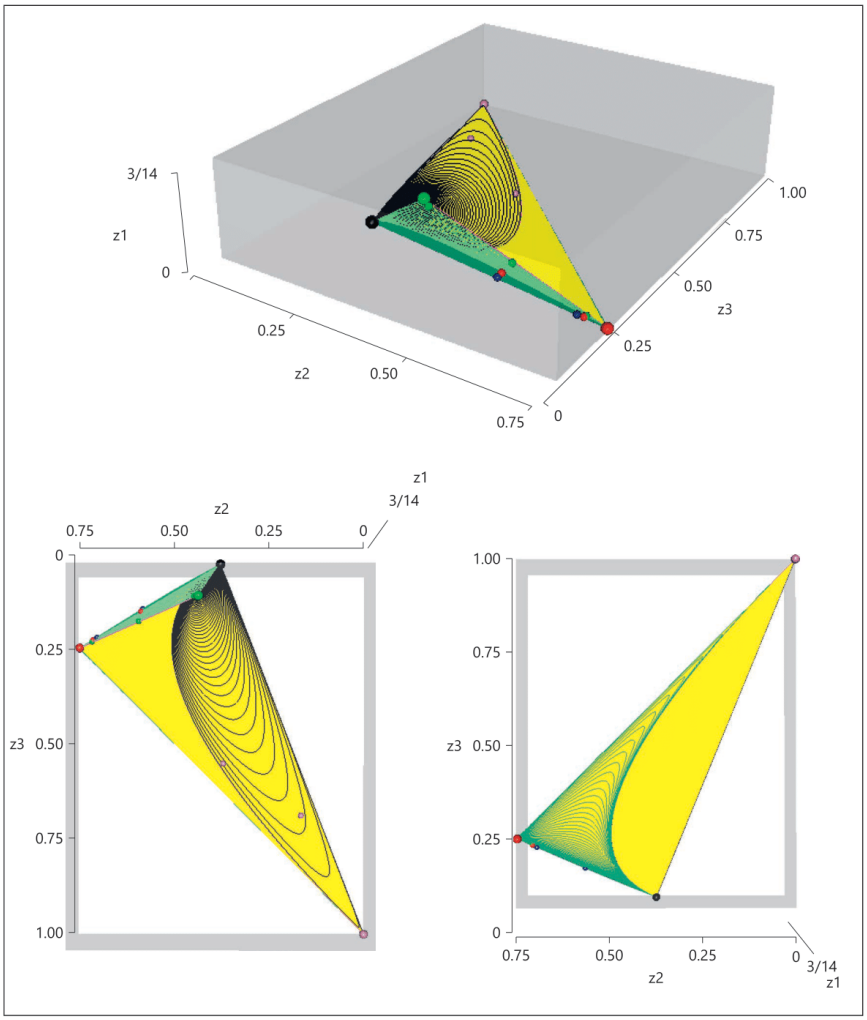
Even if the conditions under which the MOD score provides unbiased estimates of the joint trait-marker inheritance parameters are fulfilled, the identifiability of these parameters is restricted by the type(s) of pedigrees in a given sample. In a model-based linkage analysis, such as a MOD score analysis, the penetrances, disease allele frequency, and the recombination fraction represent a reparametrization of the truly underlying allele-sharing classes [9, 10, 25, 26]. In other words, allele-sharing probabilities (classes) of a given pedigree type can be expressed in terms of the joint trait-marker inheritance parameters. In the case of an affected sib pair (ASP), these allele-sharing classes are z_0, z_1, and z_2 that an ASP shares 0, 1, or 2 allele(s) IBD with restrictions to genetically possible models [27]. With $z_2 = 1 - z_0 - z_1$ and restrictions $z_1 \leq 0.5$ and $2 \times z_0 \leq z_1$, the allele-sharing classes of ASPs form a 2-dimensional parameter space – the so-called "possible triangle" [27]. Hence, as there are only 3 – 1 = 2 free parameters that can be estimated from ASP data, there will be many sets of f_0, f_1, f_2, p, and the recombination fraction θ that correspond to the estimated $\hat{z}_0$, $\hat{z}_1$, and $\hat{z}_2$. With larger pedigrees, and hence more allele-sharing classes, the degree to which the trait-model parameters can be correctly determined should be higher. However, the corresponding allele-sharing configurations have hitherto only been formulated for unilineal, affected relative pairs (e.g., affected half-sib pairs [AHSPs] [10]), ASPs [27], and affected sib triplets (ASTs) [28]. The parameter space for AHSPs is degenerated to a single line [10]. Hence, many different sets of trait-model parameters correspond to the same point on this so-called "possible line."

Using the formulas in Knapp [28], it is possible to draw the 3-dimensional parameter space for ASTs with empirically assessed restrictions for genetically possible models (Fig. 1). However, the parameter restrictions have not been derived in closed form so far. The parameter spaces for

larger pedigrees involve a larger number of dimensions, and the corresponding restrictions for genetically possible models are expected to have an even more complicated form [10, 28]. It is of note that for any type of affecteds-only analysis, the absolute values of penetrances cannot be determined, because multiplication of all penetrances by the same factor does not change the result. However, their ratios are not defined if the penetrance in the denominator of the ratio is estimated to be 0. Additionally, the ratio is subject to the estimation variance of both the penetrance in the numerator and in the denominator.

Imprinting

Genomic imprinting implies dependence of an individual's liability to develop a disease on the parental origin of the mutated allele(s), leads to a deviation from the



1

(For legend see next page.)

Hum Hered 2016;82:103–139
DOI: 10.1159/000479738

classic Mendelian assumption of equal contribution of parental genomes to the progeny and is, therefore, called a "parent-of-origin effect" [29]. In the context of a parametric linkage analysis, imprinting can be modelled using a 4-penetrance formulation distinguishing the heterozygotes according to the parental origin of the disease allele: $f = (f_0, f_{1,pat}, f_{1,mat}, f_2)$, as implemented in the program GENEHUNTER-MODSCORE (GHM) [25, 30–32], which is a further development of GENEHUNTER-IMPRINTING [33]. In the nonparametric context, the allele-sharing class z_1 of an ASP is split up into $z_{1,pat}$ and $z_{1,mat}$ according to the parental origin of the shared allele. The corresponding parameter space of ASPs, hence, extends to a 3-dimensional tetrahedron which accounts for disease models with $z_{1,pat} \neq z_{1,mat}$, i.e., for imprinting [34]. In the case of AHSPs, the allele-sharing class z_1 is distinguished as either being $z_{1,pat}$ or $z_{1,mat}$, depending on the sex of the common parent, i.e., male or female, respectively. Although the information contained in AHSPs on all trait-model parameters is limited, the information for imprinting may be high, such that parameter estimates for $f_{1,pat}$ and $f_{1,mat}$ using a sample of AHSPs having a common father and of AHSPs having a common mother should indicate imprinting if it was really present. In the case of an informative marker, this even holds if parental genotypes are missing.

In contrast, imprinting information contained in ASPs with untyped parents is 0, even in the case of a fully informative marker, because alleles shared IBD through the father cannot be distinguished from those shared IBD through the mother. However, we hypothesize that the information on linkage and imprinting gained from AHSPs can be combined with the pure linkage information contained in ASPs in the analysis to compensate for missing parental marker genotypes. If there is sufficient evidence for linkage, this pedigree scenario should lead to trait-model parameter estimates reflecting at least some degree of imprinting. Using GHM, imprinting can be quantified by looking at the imprinting index I [35], calculated from the estimated penetrances. The imprinting index equals the difference between the 2 heterozygote penetrances, normalized by the difference of the homozygote penetrances in order to properly take the case of a non-0 phenocopy rate or reduced penetrance into account:

$$I = \frac{f_{1,pat} - f_{1,mat}}{f_2 - f_0}.$$

An imprinting index of $I = 1$, therefore, indicates complete maternal imprinting (cmi), whereas $I = -1$ indicates complete paternal imprinting (cpi). If penetrances are not restricted to $f_0 < f_1 < f_2$ in the analysis, the penetrances $f_{1,pat}$

Fig. 1. Graphical depiction of the allele-sharing parameter space for affected sib triplets (ASTs). The axes notations are defined as follows (see also Knapp [28]). Axis z1: allele-sharing class z_1 with range {0; 3/14}. Axis z2: z_2 with range {0; 0.75}. Axis z3: z_3 with range {0; 1}. The panels top and at the left correspond to "top view." The boundary of the parameter space, which is defined by the genetically possible models, was empirically determined by varying the trait-model parameters $\{f_0, f_1, f_2, p\}$ in the formulas given in Knapp [28]. p, disease allele frequency; f_i, penetrances, with f_i denoting the probability that an individual with i copies of the disease allele is affected by the disease. Light green, dark green, and black lines were drawn by varying p between 0 and 1. For more details, see table below. Figures were drawn using rgl: 3D Visualization Using OpenGL, R package version 0.95.1441 (2016) by Adler, Murdoch, and others.

Boundary region Specific point	Trait model type
Light green "protruding" region	$f_1 > f_0, f_2$ or $f_1 < f_0, f_2; f_0 \neq f_2$
Dark green "bottom" region	$f_2 > f_1 > f_0$ or $f_0 > f_1 > f_2; f_1^2 > f_0 f_2; f_0$ or $f_2 = 0$
Black "top" region	$f_1^2 < f_0 f_2$; $0 \leq f_0, f_2 \leq 1; f_1 = 0$
Yellow plane (reached from the "top")	$f_1^2 < f_0 f_2; f_0 f_2 \rightarrow f_1^2$
Yellow plane (reached from the "bottom")	$f_1^2 > f_0 f_2; f_0 f_2 \rightarrow f_1^2$
Large black sphere	H_0 $(z_1, z_2, z_3) = (0.1875, 0.375, 0.1875)$
Large blue sphere	(0, 0.75, 0.25) genetically strongest additive/dominant model
Large violet sphere	(0, 0, 1) genetically strongest recessive model
Large green sphere	(3/14, 3/7, 1/7) genetically strongest overdominant model
Violet spheres	Recessive models R3, with $p = 0.01$ closer to H_0
Blue spheres	Additive models A3, with $p = 0.1$ closer to H_0
Red spheres	Dominant models D3, with $p = 0.1$ closer to H_0
Green spheres	Overdominant models U3, with $p = 0.01$ near the red sphere, and U4 near the green sphere

and $f_{1,mat}$ can, therefore, be estimated to be $<f_0$ and $>f_2$. Thus, the imprinting index may exceed 1 or fall below –1. In the case of $f_0 = f_2$, the imprinting index is defined to be 0. In a work by Haghighi and Hodge [36], it was shown that asymptotically unbiased estimates of parent-of-origin effects can be obtained using a likelihood formulation for segregation analysis without including an ascertainment parameter when ascertainment is single. The same should hold true for the method by Strauch et al. [33] applied in this paper in the context of parametric linkage analysis according to the arguments given by Ginsburg et al. [1] and Malkin and Elston [19], provided that the formulation with 4 penetrances correctly reflects the genetic mechanism of genomic imprinting.

Aims of the Present Study

The aim of the present study was to evaluate how accurately penetrances, or penetrance ratios in the case of affecteds-only analyses, and the disease allele frequency of a monogenic, dichotomous trait can be estimated in a MOD score analysis. To this end, we performed a simulation study to determine the bias and variability of trait-model parameter estimation for 6 pedigree types (AHSP, ASP, AST, discordant sib triplets [DST], discordant sib quadruplets [DSQ], and 3-generation [3-G] pedigrees) and 4 types of generic models (recessive, dominant, additive, and overdominant) as well as an imprinting model. A single marker locus linked with $\theta = 0$ to the disease locus was considered. It is of note that we did not consider the estimation of the recombination fraction θ or any LD parameter in our analysis. That is because the primary focus of this paper is on the estimation of trait-model parameters, which do not include the recombination fraction. However, the recombination fraction is confounded with the trait-model parameters, especially for smaller pedigree types, like the ones considered in our study, having only a limited number of allele-sharing classes (see also "Identifiability of Inheritance Parameters" above). In addition, LD parameters cannot be estimated using GHM so far.

We avoided the problem of an additional bias due to a possible misspecification of the sampling model for the likelihood correction. This was done by designing the simulation study in a way that conditions (i)–(iv) and (vi) mentioned above to obtain asymptotically unbiased parameter estimates from a MOD score analysis were satisfied as follows (note that only one of conditions [v]–[vii] needs to be fulfilled):

i The marker was truly linked ($\theta = 0$).
ii A diallelic autosomal binary trait locus, which is usually assumed as the mode of inheritance in a MOD score analysis, was used for the simulation of pedigree data.
iii Sampling of pedigrees was marker-independent.
iv Extension of pedigrees was trait-independent.
v –
vi Ascertainment was single in the sense of Hodge and Vieland [20].
vii –

Hence, the questions we aimed to answer in our study were:

1. For each pedigree type, can the MOD score approach differentiate between the trait-model types? That is, are, for example, recessive models recognized as being recessive, irrespective of the accuracy of the individual parameter estimates?

2. How does the estimation accuracy change from ASP to AST, i.e., when adding an affected sibling?

3. How does the estimation differ between an analysis using only affecteds vs. both affecteds and unaffecteds?

4. How does the estimation accuracy change from DST to DSQ, i.e., when adding a second unaffected sibling?

5. How does the estimation accuracy change when more complex pedigrees are considered?

6. How well can imprinting be detected and estimated in a sample of AHSPs and in a mixture sample of AHSPs and ASPs when parental genotypes are missing?

The answers to these questions are summarized in the Results section.

Methods

Nomenclature

Parameters written in capital letters (P, D, F_0, $F_{1,pat}$, $F_{1,mat}$, F_1, F_2, I) denote theoretical parameters and the parameters that were used for simulation ("true" parameters). Lowercase letters (p, d, f_0, $f_{1,pat}$, $f_{1,mat}$, f_1, f_2, i) denote the parameters that were estimated from simulated data.

Data Generation

The 5 pedigree types shown in Figure 2 (top and middle row) were chosen for the simulations. We used a sample size of 500 families for each pedigree type to ensure sufficient power to detect linkage while maintaining reasonable computation times. For certain trait-model scenarios, we performed additional analyses with a sample size of 1,000 families to assess the degree by which parameter estimates are biased due to finite sample sizes. Disease and marker locus genotypes were simulated using FastSLINK [37–39]. For each pedigree-type-trait-model scenario, we simulated 1,000 replicates. Affection statuses were assumed to be unknown for all founders. Nonfounders were either affected or unaffected (Fig. 2).

Recessive, additive, dominant, and overdominant trait models were considered in the simulations. An overview of the simulated trait models is given in Table 1. Trait models were named accord-

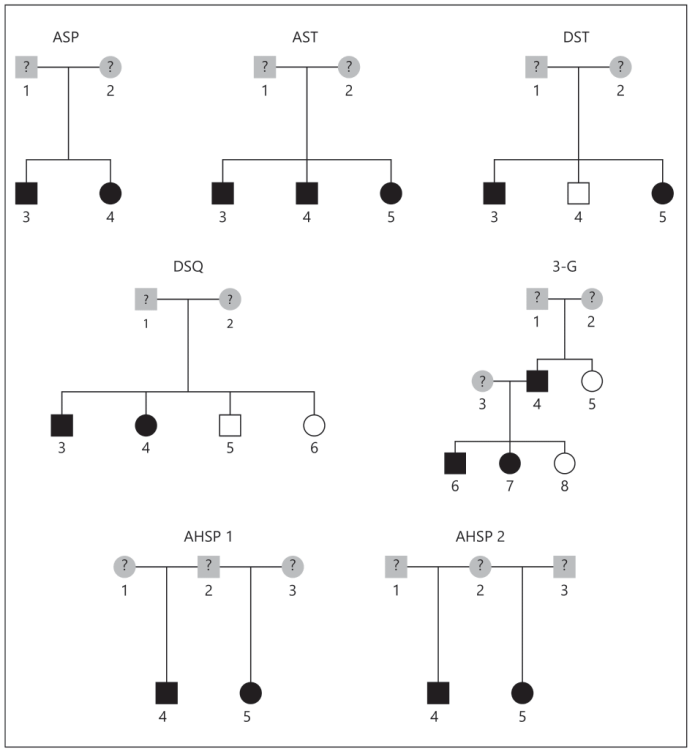


Fig. 2. Pedigree types used for the simulations. ASP, affected sib pair; AST, affected sib triplet; DST, discordant sib triplet; DSQ, discordant sib quadruplet; 3-G, three-generation pedigree; AHSP 1, affected half-sib pair with common father; AHSP 2, affected half-sib pair with common mother; ?, unknown phenotype; filled symbols, affected; empty symbols, unaffected.

ing to their generic type, i.e., "R" for a recessive model, "D" for a dominant model, "A" for an additive model, and "U" for an overdominant model. For each of the 4 generic types, 3 trait models with a particular combination of penetrances were simulated (trait-model names 1–3; Table 1). The setup of the trait-model parameters was inspired by Xing and Elston [40]. Each of the 3 trait models was simulated with a disease allele frequency $P = 0.1$ or 0.01. For the lower disease allele frequency $P = 0.01$, an additional trait model was simulated with a sample size of 1,000 families per replicate for each of the 4 generic types (Table 1). This amounts to 28 simulated scenarios. Furthermore, an overdominant model with a different combination of penetrances was simulated (model U4). For the recessive, dominant, and additive trait models, 2 further models similar to those in Flaquer and Strauch [41] were considered (models preceded by "AF" in Table 1). One of these models was simulated with sample sizes 500 and 1,000, whereas the other model was simulated with sample size 500 only. The total number of simulated scenarios, therefore, amounts to 38.

We furthermore analyzed AHSP and ASP pedigrees under a model of cpi or cmi. Differing from the scenarios in Table 1, samples contained a mixture of 2 pedigree types. Three scenarios were considered. In the first scenario, each replicate simulated under the cpi model contained 100 AHSPs who had a common father and 100 AHSPs who had a common mother (Fig. 2, bottom row). In the second scenario, each replicate simulated under the cpi model contained 100 AHSPs who had a common mother and 100 ASPs (Fig. 2, bottom row). In the third scenario, 20 AHSPs who had a common mother and 180 ASPs were simulated under the cmi model. Again, 1,000 replicates were simulated for each scenario (see Table 2 for an overview of the imprinting simulations). Imprinting was simulated using the SLINK extension SLINK Imprinting [42].

For the imprinting model, all founder genotypes were removed after data generation. The rationale behind this approach is the following: if the founder genotypes of AHSPs and ASPs are unknown, information about imprinting can only be inferred from AHSPs, with ASPs contributing only information about linkage. As a reference for comparison, a corresponding no imprinting (ni) model was considered.

Data Analysis

We used GHM version 3.1 [25] for MOD score calculation and trait-model parameter estimation. In particular, we used the GHM

Table 1. Overview of the simulated scenarios using trait models of the generic types "recessive," "dominant," "additive," and "overdominant"

Model type	Name	P	F_0	F_1	F_2	Sample size
Recessive	R1	0.01; 0.1	0.01	0.01	0.2	500 and 1,000; 500
	R2	0.01; 0.1	0.01	0.01	0.5	500
	R3	0.01; 0.1	0.01	0.01	0.8	500
	AFR1	0.2	0.04	0.04	0.2	500 and 1,000
	AFR2	0.25	0.003	0.05	0.5	500
Dominant	D1	0.01; 0.1	0.01	0.2	0.2	500 and 1,000; 500
	D2	0.01; 0.1	0.01	0.5	0.5	500
	D3	0.01; 0.1	0.01	0.8	0.8	500
	AFD1	0.05	0.04	0.2	0.2	500 and 1,000
	AFD2	0.25	0.003	0.5	0.5	500
Additive	A1	0.01; 0.1	0.01	0.1	0.2	500 and 1,000; 500
	A2	0.01; 0.1	0.01	0.2	0.5	500
	A3	0.01; 0.1	0.01	0.5	0.8	500
	AFA1	0.1	0.03	0.13	0.23	500 and 1,000
	AFA2	0.5	0.003	0.25	0.5	500
Overdominant	U1	0.01; 0.1	0.01	0.2	0.01	500
	U2	0.01; 0.1	0.01	0.5	0.01	500
	U3	0.01; 0.1	0.01	0.8	0.01	500 and 1,000; 500
	U4	0.35	0.01	0.9	0.01	500

P, disease allele frequency; F_0, F_1, F_2, penetrances with F_i denoting the probability that an individual with i copies of the disease allele is affected by the disease.

Table 2. Overview of simulated trait models with imprinting and corresponding no imprinting model

	Pedigree structure	Model name	P	F_0	$F_{1,pat}$	$F_{1,mat}$	F_2
Model with imprinting	1. 100 AHSPs with a common father + 100 AHSPs with a common mother	cpi	0.01	0	0	1	1
	2. 100 AHSPs with a common mother + 100 ASPs	cpi	0.01	0	0	1	1
	3. 20 AHSPs with a common mother + 180 ASPs	cmi	0.01	0	1	0	1
Comparison model	All structures	ni	0.01	0	0.5	0.5	1

AHSP, affected half-sib pair; ASP, affected sib pair; P, disease allele frequency; F_0, F_1, F_2, penetrances with F_i denoting the probability that an individual with i copies of the disease allele is affected by the disease; $F_{1,pat}$, $F_{1,mat}$, heterozygote penetrances distinguished by the parental origin of the disease allele (*pat*: paternally inherited, *mat*: maternally inherited); cpi, complete paternal imprinting; cmi, complete maternal imprinting; ni, no imprinting.

options "modcalc single," "penetrance restriction off," "allfreq restriction off," "maximization dense," and "dimensions 4" or "dimensions 5" for ni models and imprinting models, respectively. "modcalc single" enables a separate maximization for each genetic position. "penetrance restriction off" allows for over- and underdominant models, i.e., allows heterozygote penetrance(s) to be varied freely between 0 and 1 during the maximization. This also affects the dominance index, which is defined as

$$D = \frac{F_{1,\,pat} + F_{1,\,mat} - F_0 - F_2}{F_2 - F_0}.$$

$D = 1$ indicates a fully dominant model, whereas $D = -1$ indicates a fully recessive model. However, if the penetrances are not restricted to $F_0 < F_1 < F_2$, the dominance index may also exceed 1 or fall below –1. Note that the dominance index is defined to be 0 for models with $F_2 = F_0$, i.e., strictly overdominant or strictly underdominant models. "allfreq restriction off" allows the disease allele frequency to be estimated >0.5. "maximization dense" indicates that the MOD score is calculated for a greater number of predefined models before the fine maximization than in the standard setting. "dimensions 4" or "dimensions 5" allows all parameters (disease allele frequency plus 3 penetrances in the case of ni models or disease allele frequency plus 4 penetrances in the case of imprinting models) to be varied simultaneously in the maximization. For the models with imprinting, we ran 2 analyses. For the first, "imprinting" was set to "off" and "dimensions" to "4," and for the second, they were set to "on" and "5," to obtain ni and imprinting MOD scores, respectively. Estimates of trait-model parameters were obtained from the model yielding the highest MOD score in the analysis.

Results

Estimated values of the trait-model parameters of each simulation scenario are reported as medians based on 1,000 replicates. Sometimes, penetrances of a given replicate were estimated to be exactly 0, rendering penetrance ratios undetermined. In this case, penetrance ratios were either set to a very large number (10^6) or to 1, in case both the numerator and the denominator of the penetrance ratio were 0. Hence, no information for the estimation of the median was lost. To facilitate the comparison of the quality of estimation across pedigree types, we constructed graphics that display all 5 pedigree types using various trait models. Bias was defined as the deviation of the median estimate of a parameter from its expected value. The corresponding measure of variability is the median absolute deviation (MAD). In general, a good estimation shows both small bias and MAD (high efficiency). Impact of bias can be considered of minor importance when MAD is high. In addition to absolute penetrances, the corresponding evaluation of bias and MAD of penetrance ratios for ASPs and ASTs will be given in a dedicated section. MOD scores for each model and pedigree type are displayed in Table 3. Parameter estimation result tables for each model and pedigree type can be found in the Appendix.

Recessive Models

The parameter estimation results for recessive models can be found in Figure 3 and Appendix Tables A1, A5, and A9. With regard to recessive models, bias and MAD were often higher for ASPs compared to ASTs (Fig. 3). This is due to the fact that only 2 out of 4 parameters (3 penetrances and the disease allele frequency) are identifiable. With ASTs, 3 out of 4 parameters should be identifiable. It is of note that it is impossible in the case of affecteds-only analysis to estimate absolute penetrance values; here, only penetrance ratios, which correspond to genotype relative risks, are identifiable in the best case. Consider, for example, the 2 sets of penetrances resulting in the same MOD score: $f_0, f_1, f_2 = 0.1, 0.5, 1$ and $f_0, f_1, f_2 = 10^{-3}, 0.005, 0.01$, with the first set being more likely to be evaluated in the analysis due to the predefined trait models initially tested by GHM before the fine maximization. Generally, for all types of models (recessive, dominant, additive, and overdominant), higher MOD scores were obtained for ASTs compared to ASPs.

With ASPs, most recessive models were recognized as such, indicated by a median dominance index $d < 0$. Only R1, a model with an extremely reduced penetrance, was estimated as being additive (median $d = 0$) for $P = 0.01$. This is due to the fact that affected persons are more likely to be phenocopies in the context of a strongly reduced penetrance and a small disease allele frequency, which reduces the amount of allele sharing among affected siblings. An equivalent explanation for this can be found in Figure 4, which shows the projection of the estimated trait-model parameters for ASP pedigrees on the triangular parameter space as described in the Introduction (subsection "Identifiability of Inheritance Parameters"). For all models, the estimated values scattered around the true values without systematic deviation. However, the true value for model R1 with $P = 0.01$ lies close to the point of no linkage in the upper right corner of the triangle. In the proximity of this point, all types of generic models (recessive, dominant, additive, and overdominant) accumulate and are hard to distinguish from each other.

For ASTs, all recessive models were clearly recognized as such. Estimation accuracy of the dominance index D improved from ASPs to ASTs for most recessive models. Intriguingly, ASTs even showed the best parameter estimation performance in terms of small bias and MAD

Table 3. MOD scores of the simulated trait-model scenarios for all pedigree types

Model name	*P*	Estimated median MOD (MAD)				
		ASP	AST	DST	DSQ	3-G
R1	0.1	33.07 (5.51)	134.47 (11.19)	32.41 (5.32)	31.25 (5.36)	37.85 (5.45)
R1	0.01	0.07 (0.11)	2.89 (1.69)	0.23 (0.32)	0.27 (0.35)	0.49 (0.49)
R1n1000	0.01	0.09 (0.14)	5.58 (2.31)	0.26 (0.36)	0.31 (0.38)	0.52 (0.5)
R2	0.1	100.76 (8.31)	207.56 (12.31)	108.46 (9.06)	114.38 (9.76)	89.51 (7.01)
R2	0.01	1.08 (0.97)	116.89 (11.4)	1.15 (0.88)	1.0 (0.88)	1.84 (1.1)
R3	0.1	127.37 (9.16)	225.05 (14.13)	162.37 (10.9)	189.64 (11.69)	158.31 (9.82)
R3	0.01	5.37 (2.29)	283.25 (15.7)	4.76 (2.03)	3.86 (1.87)	8.35 (2.72)
AFR1	0.2	2.69 (1.53)	13.02 (3.53)	2.81 (1.49)	2.71 (1.52)	3.78 (1.69)
AFR1n1000	0.2	5.26 (2.19)	25.87 (5.26)	5.13 (2.07)	5.08 (1.95)	7.08 (2.39)
AFR2	0.25	40.43 (5.97)	62.95 (7.65)	50.02 (6.3)	59.07 (6.24)	33.99 (4.87)
D1	0.1	19.28 (3.72)	36.43 (5.36)	20.64 (3.87)	21.95 (4.14)	29.17 (4.97)
D1	0.01	25.19 (4.2)	91.2 (7.44)	24.93 (4.28)	24.45 (4.18)	65.87 (6.25)
D1n1000	0.01	49.92 (5.82)	182.62 (9.84)	49.26 (5.95)	48.61 (5.91)	131.43 (8.67)
D2	0.1	24.43 (4.28)	43.74 (6.06)	34.74 (4.87)	45.41 (5.99)	56.48 (6.89)
D2	0.01	47.85 (5.07)	124.71 (7.95)	54.98 (5.39)	60.9 (6.34)	117.79 (7.59)
D3	0.1	25.98 (4.38)	45.93 (6.11)	67.63 (6.82)	115.52 (9.44)	133.19 (10.3)
D3	0.01	54.44 (5.11)	134.19 (7.84)	90.84 (6.9)	125.33 (9.12)	206.27 (10.94)
AFD1	0.05	4.04 (1.9)	15.8 (3.67)	4.19 (1.81)	4.1 (1.76)	9.0 (2.66)
AFD1n1000	0.05	7.94 (2.51)	31.42 (5.15)	7.92 (2.59)	7.85 (2.55)	17.45 (3.51)
AFD2	0.25	8.94 (2.74)	12.37 (3.22)	15.77 (3.77)	24.13 (4.46)	21.30 (4.21)
A1	0.1	13.58 (3.25)	28.6 (4.72)	14.13 (3.4)	14.37 (3.44)	15.05 (3.4)
A1	0.01	7.45 (2.48)	48.12 (6.2)	7.36 (2.44)	7.17 (2.46)	28.56 (4.57)
A1n1000	0.01	14.85 (3.48)	96.12 (8.53)	14.62 (3.25)	13.99 (3.14)	57.27 (6.14)
A2	0.1	21.84 (3.93)	40.65 (5.34)	24.22 (4.08)	26.54 (4.43)	23.93 (4.22)
A2	0.01	25.46 (4.21)	90.54 (7.48)	25.42 (4.22)	25.01 (4.17)	62.34 (6.53)
A3	0.1	24.43 (3.88)	41.74 (5.63)	37.47 (5.02)	50.58 (6.05)	56.32 (6.83)
A3	0.01	47.70 (5.18)	122.61 (8.05)	55.18 (5.67)	61.37 (6.5)	116.46 (7.99)
AFA1	0.1	3.95 (1.79)	11.81 (3.3)	4.22 (1.82)	4.15 (1.72)	5.81 (2.27)
AFA1n1000	0.1	7.69 (2.43)	23.10 (4.54)	8.12 (2.42)	8.06 (2.42)	10.96 (3.12)
AFA2	0.5	2.33 (1.39)	3.28 (1.63)	3.74 (1.73)	5.38 (2.16)	3.7 (1.69)
U1	0.1	21.22 (3.97)	46.18 (5.88)	22.59 (3.98)	23.91 (4.33)	51.61 (5.54)
U1	0.01	25.33 (4.09)	93.75 (7.37)	24.99 (4.09)	24.53 (4.17)	69.53 (6.35)
U2	0.1	27.37 (4.3)	56.77 (6.8)	36.69 (5.19)	46.41 (6.11)	88.48 (6.96)
U2	0.01	48.26 (5.15)	128.75 (7.97)	55.03 (5.79)	60.94 (6.36)	123.23 (7.73)
U3	0.1	29.17 (4.32)	59.87 (6.86)	68.25 (6.92)	109.36 (8.95)	153.82 (9.67)
U3	0.01	54.91 (4.99)	139.16 (7.7)	90.54 (6.97)	124.03 (9.13)	208.56 (10.64)
U3n1000	0.01	109.07 (6.81)	277.17 (11.52)	180.17 (10.03)	247.99 (13.17)	416.58 (15.56)
U4	0.35	12.4 (3.3)	19.99 (4.24)	66.74 (6.53)	136.54 (9.11)	149.85 (8.58)

MAD, median absolute deviation, adjusted by a constant (1.4826) for asymptotically normal consistency; ASP, affected sib pair; AST, affected sib triplet; DST, discordant sib triplet; DSQ, discordant sib quadruplet; 3-G, 3-generation pedigree; *P*, true value for the disease allele frequency.

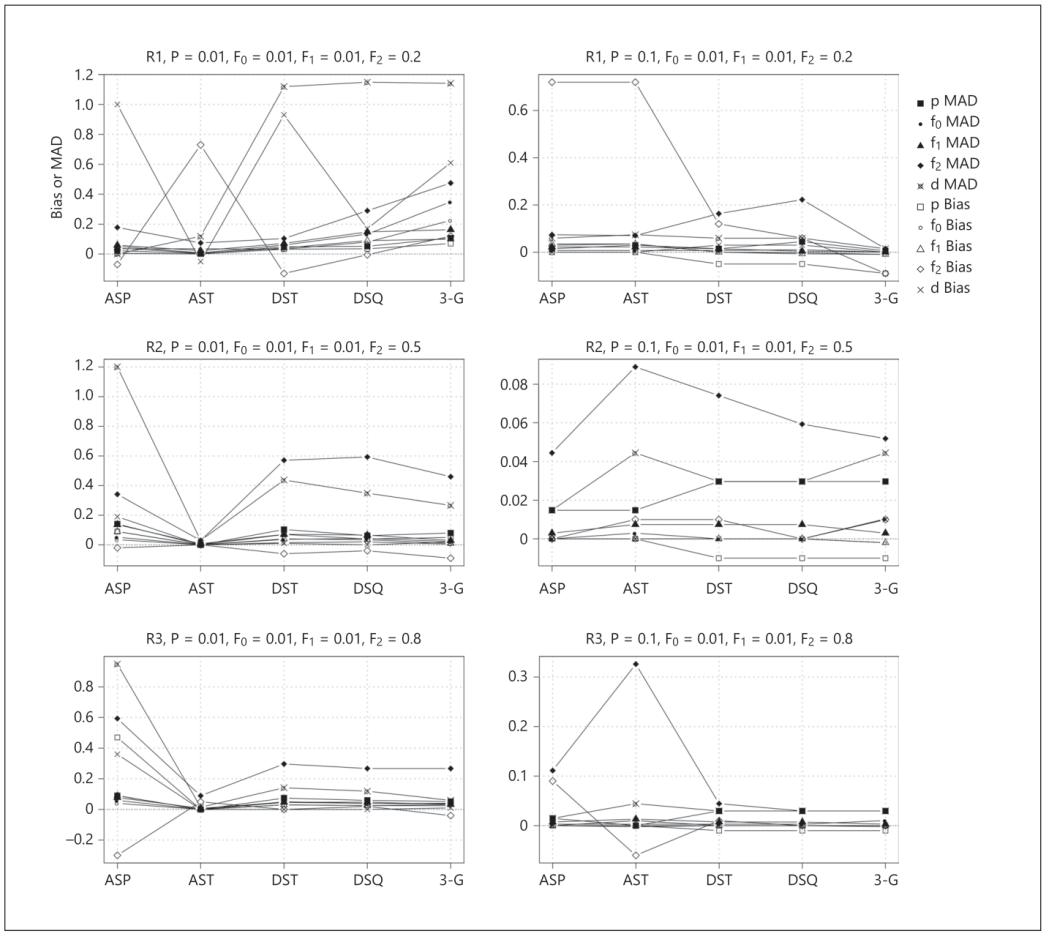
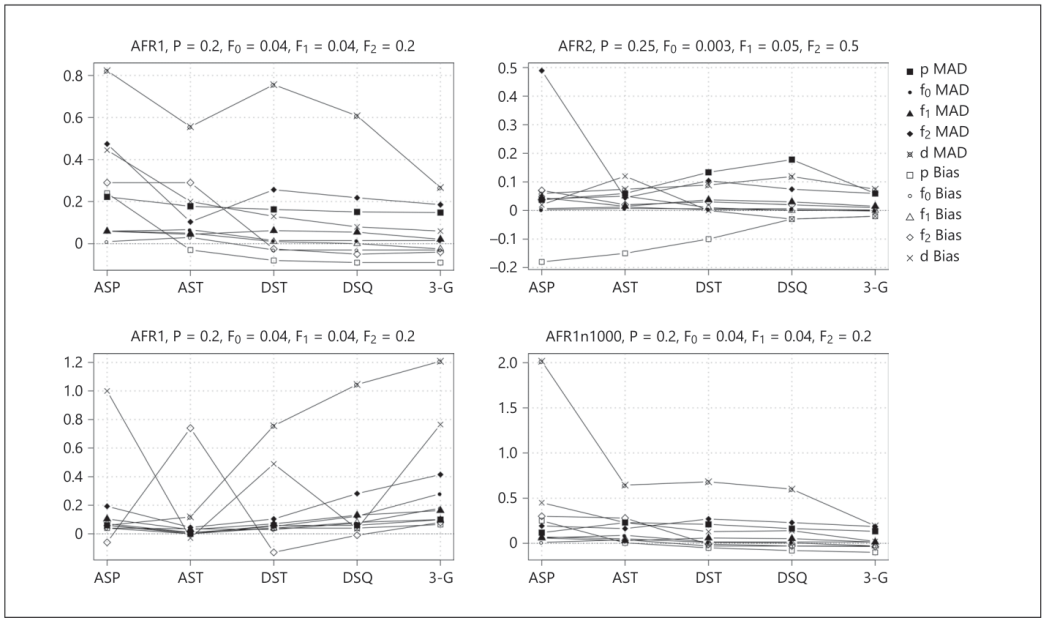


Fig. 3. Illustration of bias and variability of the parameter estimation for recessive models using different pedigree types. The trait-model parameters used for the simulations are given above the panels for each trait model. Estimations of the individual parameters are depicted by five unique symbols. For each parameter, the median absolute deviation (MAD) and bias, defined as *bias* = *median* (true parameter value – estimated value), are plotted. Pedigree types (for details, see Fig. 2 and its legend) are displayed on the *x*-axis with increasing complexity, i.e., ASPs are located on the very left side and 3-G pedigrees are located on the very right side. *p*, disease allele frequency; f_i, penetrances; *d*, dominance index.

(Figure continued on next page.)

across all investigated pedigrees for models R2 and R3 both with $P = 0.01$ (Fig. 3). This might be explained as follows: although only penetrance ratios can in theory be estimated using ASTs, the corresponding set of absolute values of the penetrances resulting in such high ratios (F_2/F_1: 50 for R2 and 80 for R3) is limited in a maximization starting with a fixed grid of genetically plausible values (the genotype relative risk of model R1 with $P = 0.01$ obviously was too low to show the aforementioned effect). Further, despite the small disease allele frequency, a



3

low phenocopy rate together with a high penetrance ensures enough information for the estimation of F_2 in relation to F_0 and F_1 in the context of ASTs. In addition, the number of degrees of freedom in an AST MOD score analysis is lower compared to an analysis with pedigrees containing healthy individuals, which can lead to a higher power of an affecteds-only analysis (see also Flaquer and Strauch [41]) and hence to a more efficient parameter estimation for some model types (up to a constant factor multiplied to all penetrances).

With regard to DSTs and DSQs, all models were correctly classified as being recessive, and the median dominance index was mostly close to its expected value. In most cases, median estimates of all parameter values were similar for the 2 pedigree types. When the true disease allele frequency was small ($P = 0.01$), it was always overestimated. When it was large ($P \geq 0.1$), it was always underestimated. Penetrances F_0 and F_1 were estimated with high accuracy for models R1–R3 with $P = 0.1$ and model AFR2. For models R1–R3 with $P = 0.01$, F_0 and F_1 were overestimated. In the case of model AFR1, F_0 was underestimated; however, F_1 was estimated with good accuracy. Median estimates of F_2 were close to their expected values for most models, with higher accuracy for DSQs compared to DSTs. In general, F_2 could be estimated more accurately for stronger genetic models, which is the case for the investigated recessive models with higher penetrance and disease allele frequency. MOD scores were comparable for DSTs and DSQs (Table 3), except for models R2 ($F_2 = 0.5$) and R3 ($F_2 = 0.8$) with $P = 0.1$ as well as model AFR2 ($F_2 = 0.5$). This is due to the fact that an additional healthy individual increases linkage information only if penetrance and genotype relative risk are sufficiently high ($F_2 >> F_0, F_1$ for a recessive model).

Using 3-G pedigrees, median estimated dominance indices were all close to their expected values except for model R1 with $P = 0.01$. The estimation of the disease allele frequency was accurate for models R2 and R3 both with $P = 0.1$ and AFR2 with $P = 0.25$. The median F_0 and F_1 penetrances were estimated with good accuracy for models R3 with $P = 0.1$, R2, and AFR2. The homozygous mutant penetrance F_2 was estimated with good accuracy for models R2 with $P = 0.1$, R3, and AFR2. However, in all other cases, the estimated median F_2 was still larger than the corresponding medians for F_0 and F_1.

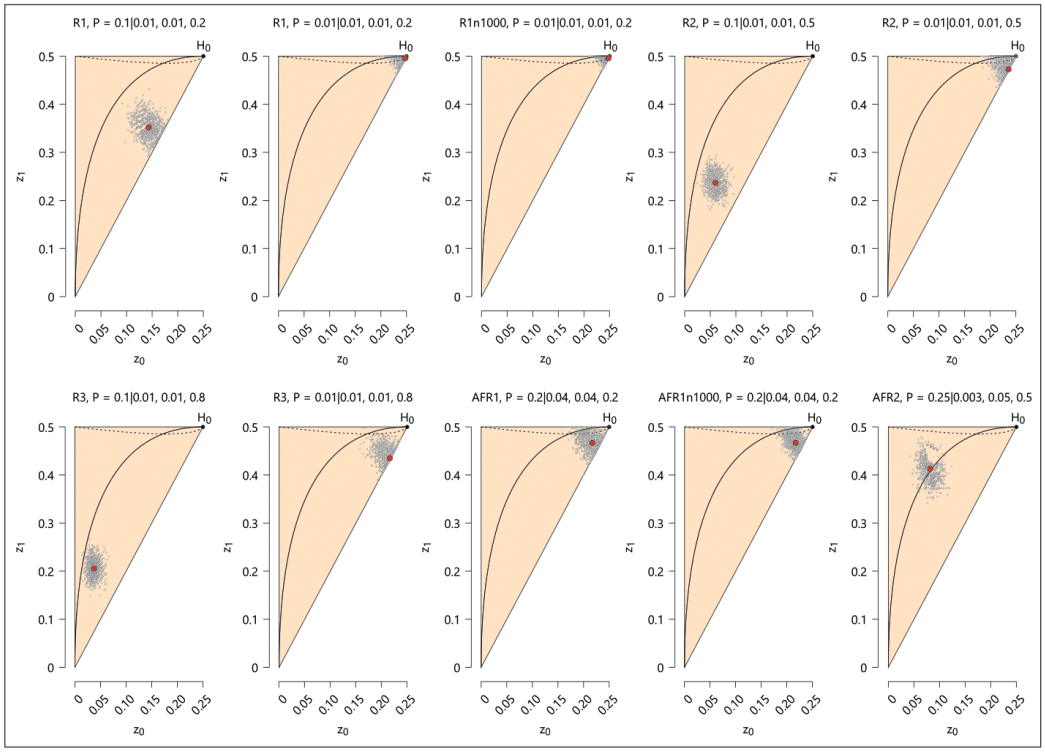


Fig. 4. Recessive models: projections of trait-model parameter estimates on the possible triangle parameter space of affected sib pairs (ASPs). The trait-model parameters used for the simulation are given above the panels for each trait model, and its projection in terms of allele-sharing is depicted by a red dot. z_0, allele-sharing probability that an ASP shares no allele identical-by-descent (IBD); z_1, allele-sharing probability that an ASP shares 1 allele IBD; trait-model parameters used for the simulation: disease allele frequency P and penetrances F_0, F_1, F_2.

With regard to DSTs, DSQs, and 3-G pedigrees, bias was often smaller than MAD across all models, yet especially large MADs were obtained for penetrance F_2 and models R2 and R3 both with $P = 0.01$ as well as for AFR1. Values for bias and MAD did not consistently decrease when moving from DSTs over DSQs to 3-G pedigrees, except for the weak genetic model R1 with $P = 0.01$ and AFR2 (Fig. 3). Better parameter identifiability when moving from ASTs to DSTs as measured by a reduction in bias, especially of the F_2 penetrance, could only be observed for models R1 with $P = 0.1$ and AFR1.

Dominant Models

Parameter estimation results for dominant models are given in Figure 5 and Appendix Tables A2, A6, and A10. The estimation of individual parameters for ASPs and ASTs was not very accurate, which is in line with the fact that exact penetrances cannot be estimated for affecteds-only pedigrees, as explained above. The median dominance index was underestimated for all models, some of which were even misclassified as being additive. In the case of ASPs, this can be explained by the proximity of both model classes in the triangular parameter space (Fig. 6, 7). In particular, dominant models without phenocopies are represented by the dashed line, whereas ad-

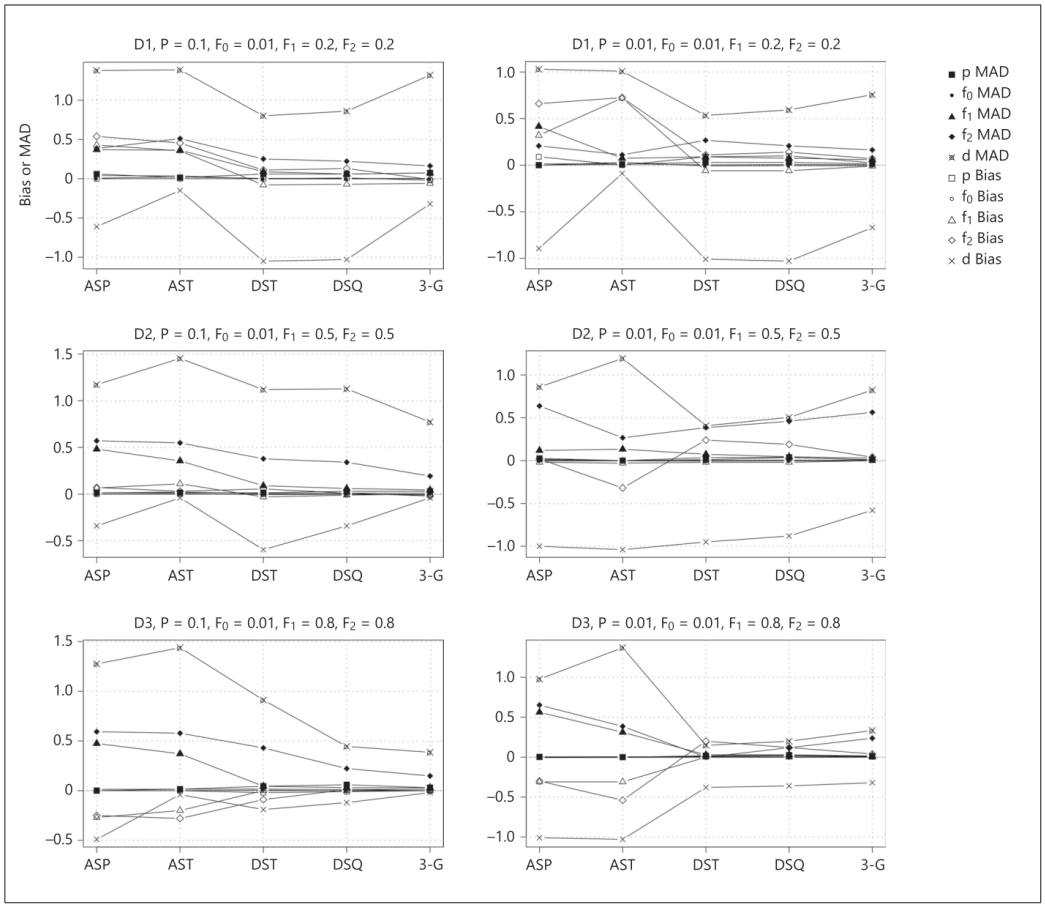
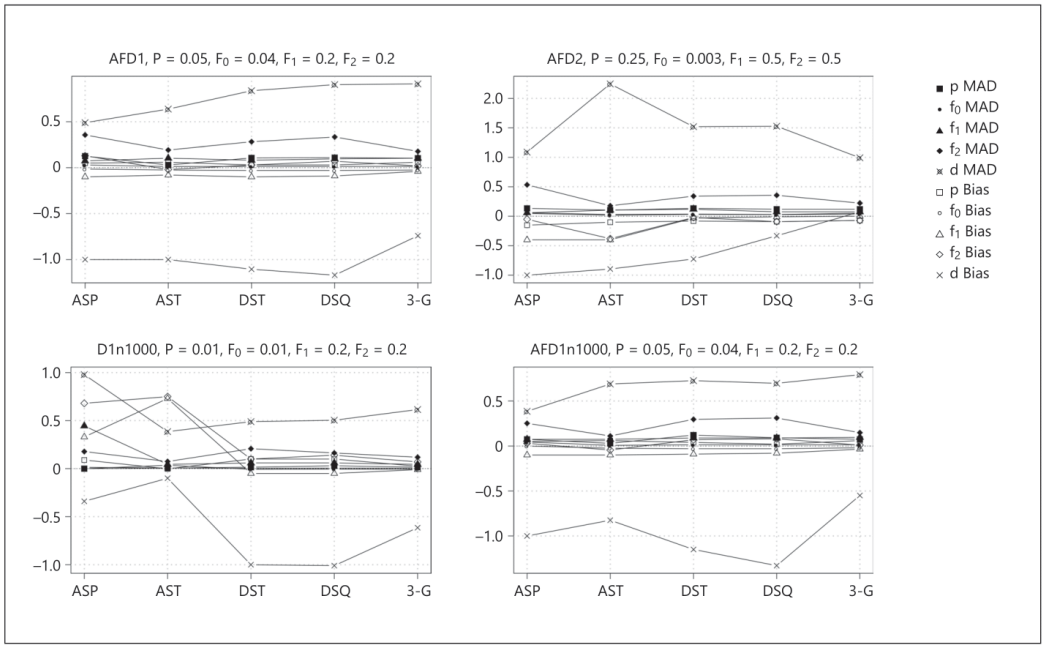


Fig. 5. Illustration of bias and variability of the parameter estimation for dominant models using different pedigree types. For more details, see Figure 3.

(Figure continued on next page.)

ditive models lie on the upper edge of the triangle. Hence, models D1–D3 with $P = 0.01$ and AFD1, which are located closest to the upper edge of the triangle, showed a median estimated dominance index d close to 0, corresponding to an additive model. The estimation of the dominance index improved when analyzing ASTs instead of ASPs for most models. The same holds for the disease allele frequency, albeit to a lesser degree.

For DSTs and DSQs, many models were misclassified as rather additive for both pedigree types when looking at their corresponding dominance indices. Only the median dominance index d for model D3 with $P = 0.1$ clearly pointed to dominance ($d = 0.81$ for DSTs and $d = 0.88$ for DSQs). Otherwise, median dominance indices for models D2, D3, and AFD2 were all positive but clearly below 1 for both pedigree types. Models D1 and AFD1 even showed median d values around 0 and below 0, re-



5

spectively. The disease allele frequency was estimated accurately for models D1–D3 with $P = 0.1$, overestimated for models D1–D3 with $P = 0.01$ and the AFD1 model, and underestimated for the AFD2 model. Estimates of P were comparable between both pedigree types. Penetrance F_0 was mostly underestimated for models D1–D3 and AFD1 using both pedigree types. With regard to F_1, models D2, D3, and AFD2 showed good accuracy for both pedigree types, whereas it was underestimated for models D1 and AFD1. F_2 was often overestimated. Similar to recessive models, MOD scores were comparable between DSTs and DSQs (Table 3), except for models D2 (F_1, $F_2 = 0.5$), D3 (F_1, $F_2 = 0.8$), and AFD2 (F_1, $F_2 = 0.5$). As before, this is due to the fact that an additional healthy individual increases linkage information only if penetrance and genotype relative risk are sufficiently high (F_1, $F_2 >> F_0$ for a dominant model). Only in this case, penetrance estimation is also improved for DSQs compared to DSTs.

In the case of 3-G pedigrees, median d values pointed towards dominance for all models. Median dominance indices were close to their expected values for models D2 and D3 with $P = 0.1$ as well as model AFD2. Estimates of the disease allele frequency showed good accuracy for models D1 with $P = 0.1$, D2, and D3. Estimates for F_0 were mostly close to the expected value. Estimates for F_1 and F_2 were very close to their expected values, with the highest accuracy for models D2 and D3.

With respect to dominant models, bias and MAD decreased when moving from ASPs over ASTs, DSTs, and DSQs to 3-G pedigrees for models D1, D2, and D3 all with $P = 0.1$ (Fig. 5). Median bias for F_2 seemed to be unduly small for ASPs for model D2 with $P = 0.01$. This can be explained by looking at the corresponding parameter distribution for ASPs (data not shown), which showed that F_2 was mostly estimated near 0 (<0.1 in 25.3% of the replicates) or 1 (>0.9 in 36.6% of the replicates). This is also reflected in the high MAD of F_2 (Fig. 5). Generally, for all dominant models, bias and MAD mostly decreased when moving from affecteds-only pedigrees over DSTs and DSQs to 3-G pedigrees. Only for model AFD1, the results were similar across all pedigree types. Bias was mostly smaller than MAD across all models for DSTs, DSQs, and 3-G pedigrees.

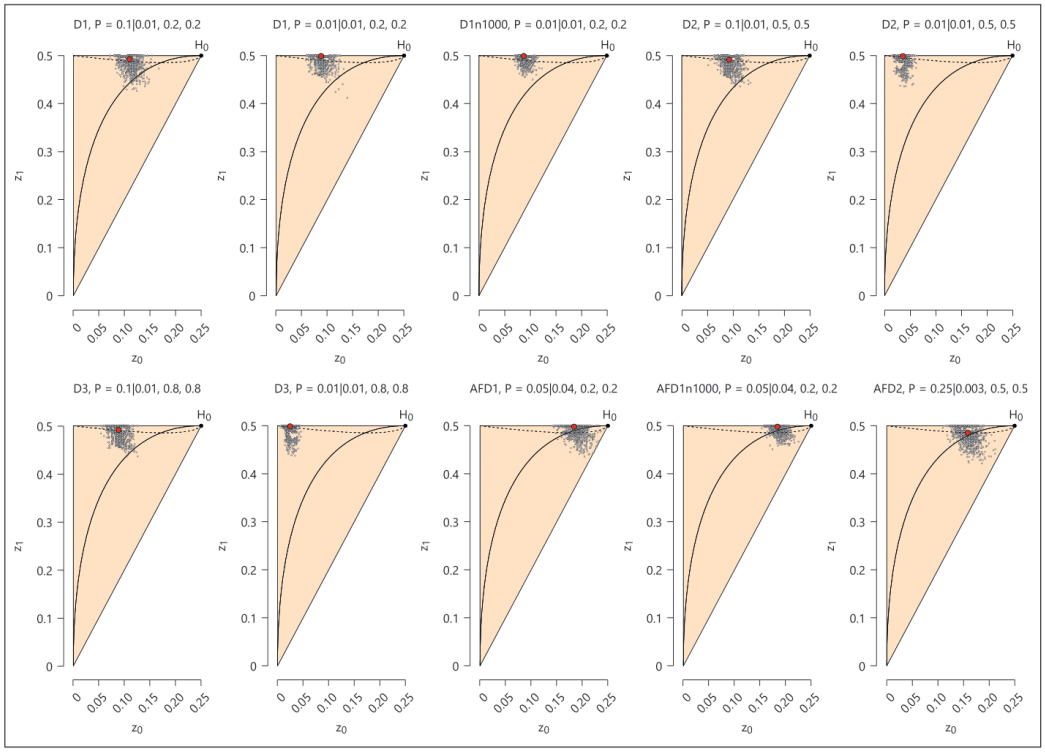


Fig. 6. Dominant models: projections of trait-model parameter estimates on the possible triangle parameter space of ASPs. For more details, see Figure 4.

Additive Models

Parameter estimation results for additive models are depicted in Figure 8 and Appendix Tables A3, A7, and A11. For ASPs, the projection of estimated trait-model parameters on the triangular parameter space, as displayed in Figure 7, illustrates that all additive models are very close to the upper edge of the triangle. Model AFA2, which has the weakest genetic effect among the investigated additive models, shows the largest distance to strictly dominant models (dashed line in Figure 7) within the allele-sharing parameter space of ASPs. The median estimated dominance indices d were close to their expected values for both ASPs and ASTs, except for model A2, which showed deviation towards dominance, and model A3. For most models and both pedigree types, the median estimated disease allele frequency p was also close to the expected value. Again, the estimation of individual penetrances for ASPs and ASTs was not very accurate, given that these pedigree types contain only affected individuals.

For DSTs and DSQs, the median dominance indices tended towards their expected values, but were not accurate for most models. The estimation of the disease allele frequency was comparable between DSTs and DSQs and showed good accuracy for models A1 with $P = 0.01$ as well as models A2 and A3 both with $P = 0.1$. Otherwise, models with $P = 0.01$ showed an overestimated disease allele frequency (A2, A3), whereas for models with $P \geq 0.1$ it was underestimated (A1, AFA1, AFA2). Penetrances F_0 and F_1 were estimated accurately for all models and

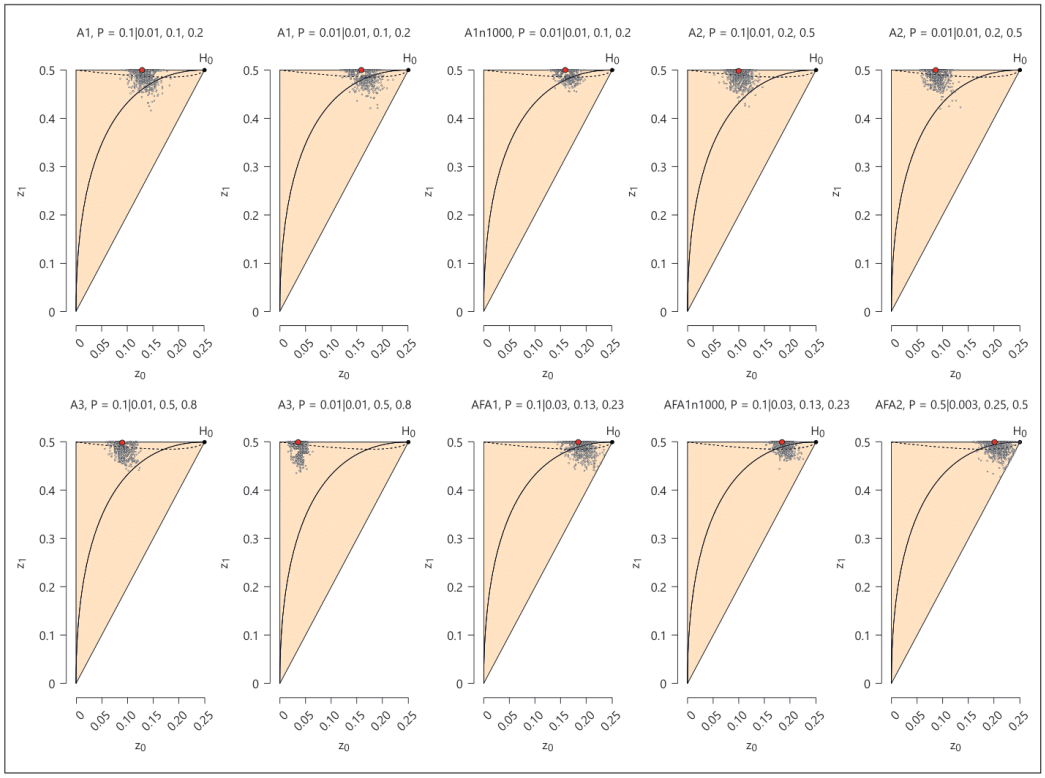


Fig. 7. Additive models: projections of trait-model parameter estimates on the possible triangle parameter space of ASPs. For more details, see Figure 4.

both pedigree types, with a slight underestimation in some cases. F_2 was estimated with acceptable accuracy for both pedigree types; however, it was always underestimated, most prominently for model A2 with $P = 0.01$ ($F_2 = 0.5$; $f_2 = 0.335$ for DSTs and $f_2 = 0.35$ for DSQs). The parameter estimation did not substantially improve when using DSQs instead of DSTs (Fig. 8). This is in line with the MOD scores in Table 3, which were comparable between DSTs and DSQs, with only a slight increase for models A3 and AFA2. As before, this is due to the fact that models A3 and AFA2 show the highest penetrance and genotype relative risk among the investigated models, such that an additional healthy individual can contribute at least some extra linkage information in the analysis.

The accuracy of median d values for additive models was not very high when using 3-G pedigrees in the analysis. However, most dominance indices still pointed to additivity. The results for the disease allele frequency showed good accuracy for models A1 and A2, each with $P = 0.1$, A3, and AFA1. The estimates for penetrance F_0 showed good accuracy for most models. Median estimates for F_1 were mostly identical to their expected value. Penetrance F_2 was estimated with good accuracy for models A1 and A2, each with $P = 0.1$, A3, AFA1, and AFA2.

The results for the additive models in Figure 8 showed a general trend towards less bias when moving from affecteds-only pedigrees over DSTs and DSQs to 3-G pedigrees, except for model AFA1. When moving from DSTs over DSQs to 3-G pedigrees, MAD slightly decreased ex-

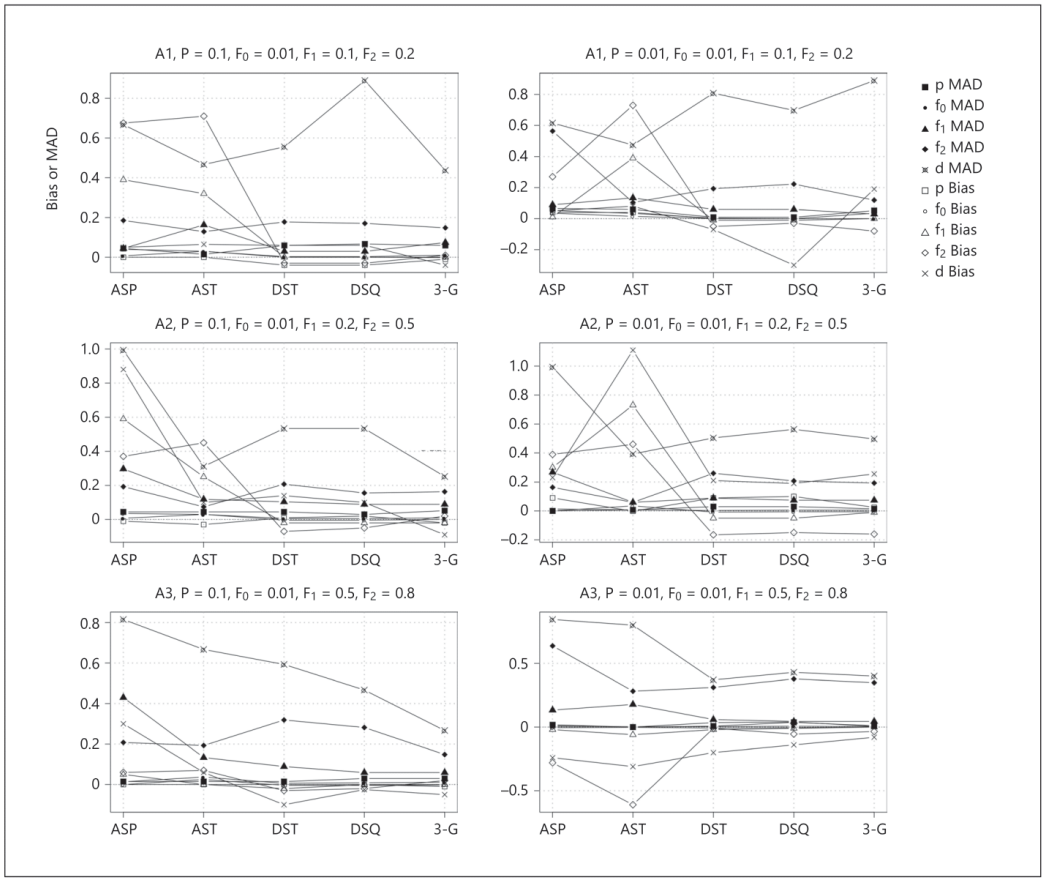


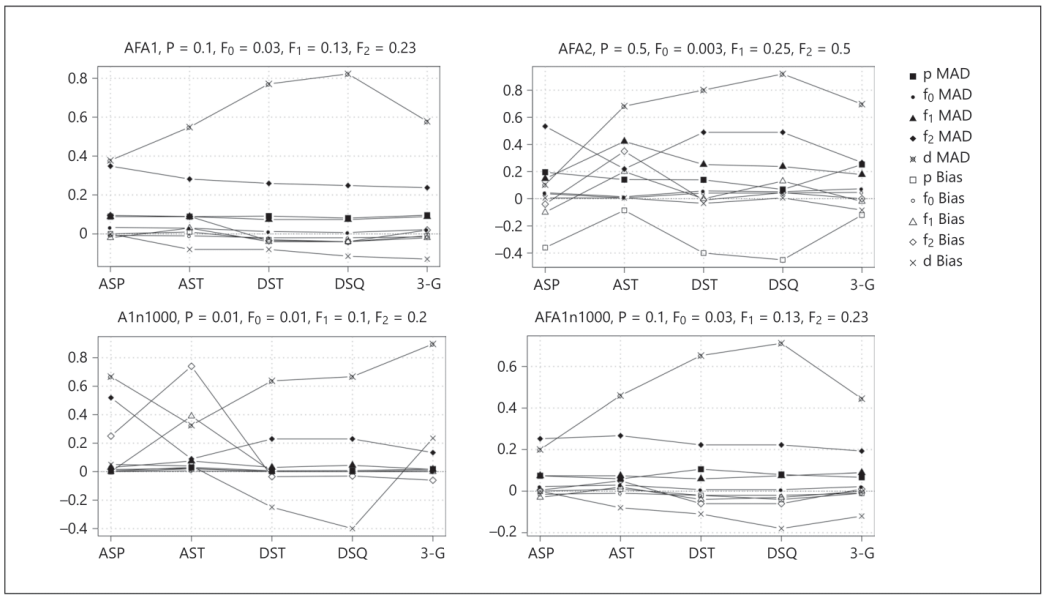
Fig. 8. Illustration of bias and variability of the parameter estimation for additive models using different pedigree types. For more details, see Figure 3.

(Figure continued on next page.)

cept for models A3 with $P = 0.01$ and AFA1. Bias was mostly smaller than MAD across all models for DSTs, DSQs, and 3-G pedigrees.

Overdominant Models

Parameter estimation results for overdominant models are given in Figure 9 and Appendix Tables A4, A8, and A12. As already mentioned above, the dominance index D is defined to be 0 for models with $F_0 = F_2$, because the denominator would be 0. Therefore, D cannot serve as a performance measure for the analyzed overdominant models. For ASPs and most models, the median disease allele frequency p was estimated close to the expected value. Overdominance, i.e., $F_0 < F_1$ and $F_2 < F_1$, was correctly assessed for models U1–U3 with $P = 0.1$ and U3 with $P = 0.01$ and a sample size of 1,000 pedigrees (Appendix Table A4). All other models were classified as rather additive (e.g., U1 with $P = 0.01$) or dominant (e.g., U2 with $P = 0.01$). The projections of the estimated trait-model parameters in the parameter space of ASPs are shown in Figure 10. The allele-sharing estimates of particular models were not evenly distributed around the true point (e.g.,



8

U2 with $P = 0.1$ and U3 with $P = 0.1$), which might be due to peculiarities of the parameter space. The true point for model U1 with $P = 0.01$ lies between the upper edge of the triangle, which corresponds to additive models, and the dashed line, representing dominant models. The location and distribution of the estimates for this model resembled those of the additive model A3 with $P = 0.1$ depicted in Figure 7. Indeed, the median estimates for U1 with $P = 0.01$ and A3 with $P = 0.1$ were similar for all penetrances as well as the disease allele frequency.

With regard to ASTs, the median disease allele frequency p was estimated close to the true value for all models. Overdominant models could be better distinguished from other model types when using ASTs instead of ASPs, because the corresponding allele-sharing values form a unique, separated compartment of the 3-dimensional parameter space (Fig. 1). Hence, overdominance was correctly assessed for all models except model U4 (Fig. 9, ASTs). Why model U4 was so difficult to be classified as overdominant for ASPs and ASTs can be explained as follows. As can be seen from Figures 1 (ASTs) and 10 (ASPs), model U4 occupies a distinct part of the parameter space as compared to models U1–U3. For both pedigree types, however, it can be shown that this distinct part of the parameter space can as well be occupied by underdominant models, i.e., $F_0 > F_1$ and $F_2 > F_1$, which is reflected by the corresponding median penetrance estimates for ASTs (Appendix Table A4).

For DSTs and DSQs, estimates of the disease allele frequency for models U1–U3 with $P = 0.1$, model U3 with $P = 0.01$, and model U4 showed good accuracy; otherwise, it was clearly overestimated. Median F_0 penetrances were estimated around their expected value (0.01) for both pedigrees, albeit slightly underestimated. Estimations of the median F_1 penetrance were accurate for all models, except model U1. Estimating penetrance F_2 proved to be difficult, since only models U3 with $P = 0.1$ and U4 showed values that were near their expectations. Generally, an estimation of F_2 is difficult when the disease allele frequency is low, because only a few individuals of the dataset actually have a homozygous mutant genotype and can contribute information to the estimation of F_2. Therefore, the relations $F_0 < F_1$ and $F_2 < F_1$ were only identified for models U2 and U3 both with $P = 0.1$ and U4 with $P = 0.35$ for both pedigree types. As explained above, the additional healthy individual in DSQs can increase the MOD score only if the penetrance and the genotype relative risk are sufficiently high, which is the case for models U2, U3

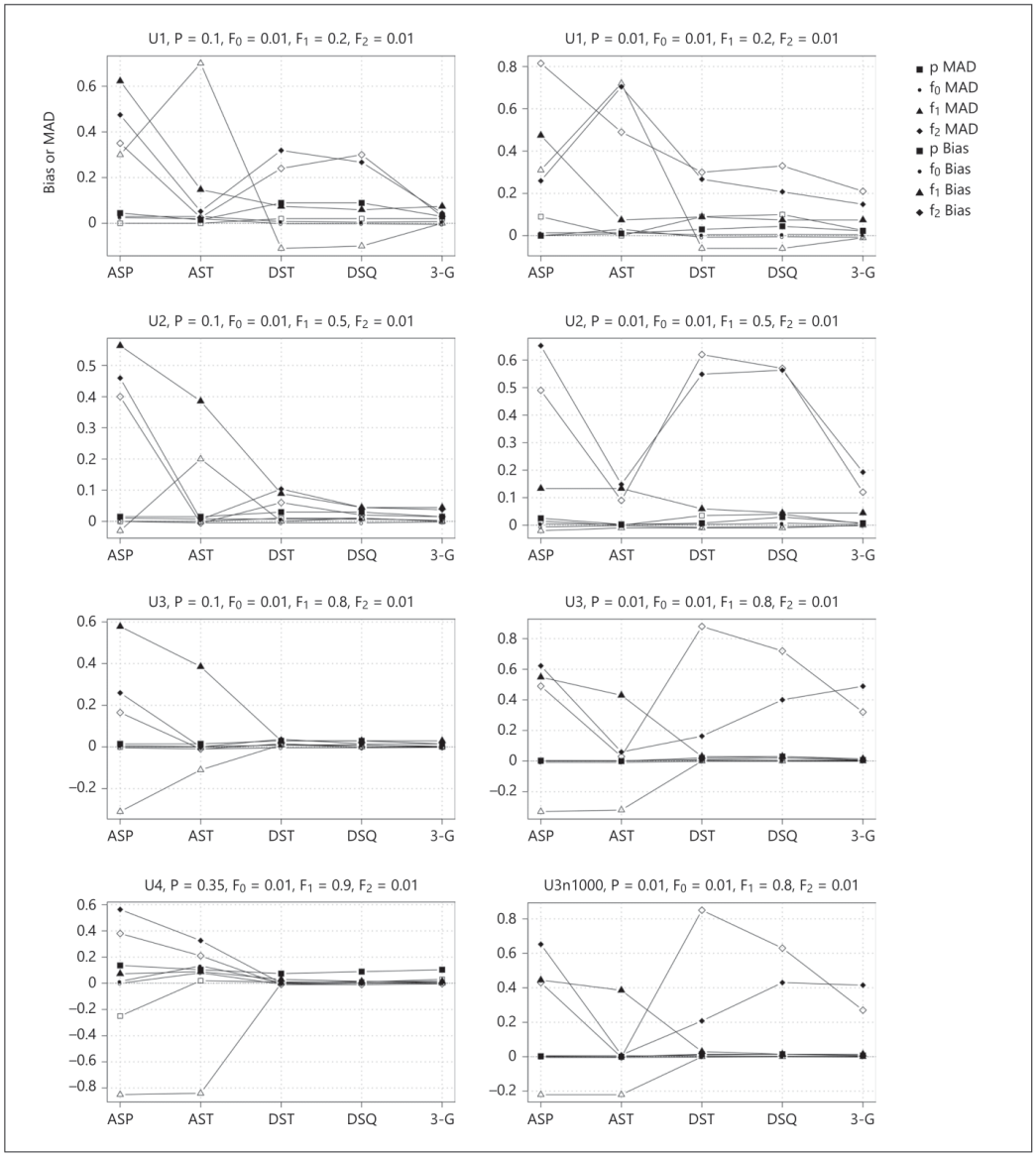


Fig. 9. Illustration of bias and variability of the parameter estimation for overdominant models using different pedigree types. For more details, see Figure 3.

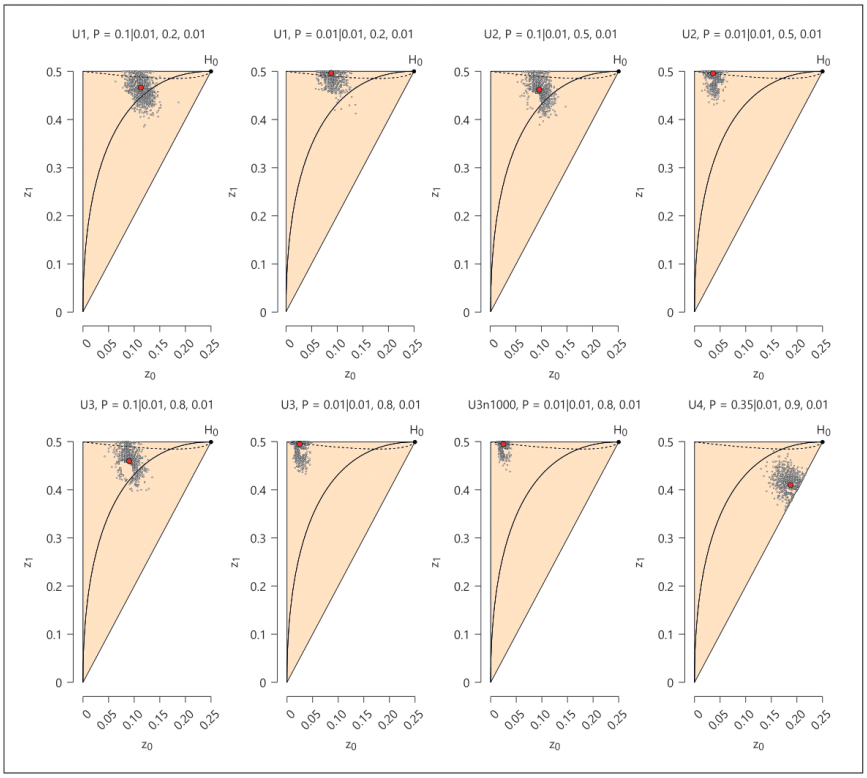


Fig. 10. Overdominant models: projections of trait-model parameter estimates on the possible triangle parameter space of ASPs. For more details, see Figure 4.

and U4 (Table 3). By the same token, when adding a second healthy individual, penetrance estimation was also improved for model U3 with $P = 0.01$, which pointed to overdominance only for DSQs but not for DSTs (Appendix Table A8).

Using 3-G pedigrees, the estimation of the disease allele frequency showed good accuracy for most models, especially for models U1–U3 with $P = 0.1$ and model U4. The median penetrances F_0 were estimated near their expected value (0.01) for all models, albeit slightly underestimated in most cases. Penetrance F_1 was estimated with very high accuracy, with all but one medians estimated exactly at the expected value. The accuracy of the estimation of F_2 depended on the disease allele frequency – models with $P \geq 0.1$ showed good accuracy, whereas F_2 was always overestimated for models with $P = 0.01$. As mentioned above, when the disease allele frequency is low, the dataset contains too few individuals with a homozygous mutant genotype that can contribute to the estimation of F_2. However, median estimates of F_2 were significantly lower than those of F_1, which clearly indicates overdominance, except for model U1 with $P = 0.01$.

For models U1–U3 with $P = 0.01$, median bias of F_2 was high, especially for ASPs, DSTs, and DSQs (Fig. 9). This is due to the fact that ASPs, DSTs, and DSQs contain only 2 affected individuals, compared to ASTs and 3-G pedigrees having 3 affected individuals. The additional affected individual results in a larger number of mutant alleles per pedigree and hence in a larger number of ho-

mozygous mutant individuals. Bias and MAD decreased when moving from DSTs over DSQs to 3-G pedigrees for most models. Better identifiability of parameters as measured by a reduction in bias when using pedigrees with unaffected individuals could only be observed for models U1 with $P = 0.01$, U3 with $P = 0.1$, and U4. In DSTs, DSQs, and 3-G pedigrees, bias was often larger than MAD for F_2. As can be seen from Figure 9, parameter estimation results were best for models with $P \geq 0.1$, especially when using 3-G pedigrees.

Penetrance Ratios for ASPs and ASTs

As already mentioned above, the exact numerical values for trait-model parameters cannot be obtained from affecteds-only analyses. However, the corresponding penetrance ratios can in principle be estimated. In Table 4, we present the estimation of pairwise ratios of the penetrances F_0, F_1, F_2 for all models in our affecteds-only analyses with ASPs and ASTs. Generally, the variability (as measured by MAD in our case) for penetrance ratios is expected to be higher than for the corresponding individual penetrances, especially when the expected penetrance ratio is high.

For recessive models and ASPs, the 3 penetrance ratios (F_1/F_0; F_2/F_0; F_2/F_1) were estimated with best accuracy for models R1, R2, and R3 with the larger disease allele frequency $P = 0.1$. The ratio between F_1 and F_0, which equals 1 for all recessive models except model AFR2, was usually well recognized, whereas F_2/F_0 and F_2/F_1 were underestimated for models with $P = 0.01$. There was a clear improvement in the estimation of the penetrance ratios F_2/F_0 and F_2/F_1 when using ASTs for the models with disease allele frequency $P = 0.01$ and the AFR1 model. Only the models R1, R2, and R3, each with $P = 0.1$, as well as AFR1 showed a smaller bias than MAD for both ASPs and ASTs and for all penetrance ratios. While bias of penetrance ratios often decreased when using ASTs instead of ASPs, the corresponding MAD was often higher, especially for models with $P = 0.01$ (Table 4).

For dominant models, the penetrance ratio that was close to 1, i.e., F_2/F_1, was overestimated for ASPs, albeit only slightly for models D1–D3 with $P = 0.1$ and D1 with $P = 0.01$. The ratios F_1/F_0 and F_2/F_0 were mostly underestimated for models D1–D3 with $P = 0.1$ and AFD2, or mostly overestimated for models D1–D3 with $P = 0.01$ and AFD1. The estimation of ratios improved with ASTs compared to ASPs only for models D1–D3 with $P = 0.01$. In the case of ASPs, bias was smaller than MAD for all penetrance ratios and models, except for AFD1n1000 and AFD2. For ASTs, in addition to AFD1n1000 and AFD2, higher bias than MAD was also obtained for models D2 and D3, each with $P = 0.1$.

For additive models, penetrance ratios were estimated best for models AFA1 and A1, which were strictly additive or close to strictly additive, respectively. In general, the benefit for the accuracy of the estimation of penetrance ratios when using ASTs instead of ASPs was not as clear-cut as for the other models. Here, the estimation mostly improved for one ratio and deteriorated for another one. For ASPs and ASTs, bias was smaller than MAD for all penetrance ratios and models, except for models A2 and AFA2, and, in the case of ASTs, model A3 with $P = 0.1$.

Results for the overdominant models and ASPs showed that the penetrance ratio F_1/F_0 was underestimated for models U1 and U2 with $P = 0.1$ as well as model U4 with $P = 0.35$, and overestimated for models with $P = 0.01$. The other 2 ratios, F_2/F_0 and F_2/F_1, were always overestimated, even to a higher degree for models with $P = 0.01$. The penetrance ratios for model U4 could not be estimated accurately, for neither ASPs nor ASTs, due to the confounding of over- and underdominant models, as explained above. In most other cases, there was a clear improvement in estimation accuracy of all 3 penetrance ratios when using ASTs compared to ASPs. For both pedigree types, bias was mostly smaller than MAD for all penetrance ratios and models.

Summary of Trait-Model Parameter Estimation Results

The results are summarized as answers to questions (1)–(5) given in the Introduction section.

(1) The ability of the MOD score approach to differentiate between the trait-model types (recessive, dominant, additive, and overdominant) was limited by the underlying parameter spaces of the corresponding pedigrees in the analysis. Among the recessive models, a stronger genetic effect provided a better discrimination from other model types across all sorts of investigated pedigrees. Adding one unaffected individual to an ASP pedigree was mostly sufficient to identify and correctly estimate the parameters of the recessive model. Additive and dominant models were generally hard to discriminate using affecteds-only data due to their spatial proximity in the corresponding allele-sharing parameter space. The discrimination between additive and dominant models improved by adding unaffected individuals and when using 3-G pedigrees. The correct classification of overdominant models substantially improved from ASPs to ASTs. With 3-G pedigrees, trait-model parameters of overdominant

Table 4. Estimation of penetrance ratios for ASPs and ASTs

Model name	P	F_1/F_0	Estimated median f_1/f_0 (MAD)		F_2/F_0	Estimated median f_2/f_0 (MAD)		F_2/F_1	Estimated median f_2/f_1 (MAD)	
			ASP	AST		ASP	AST		ASP	AST
R1	0.1	1	1.0 (0.74)	1.0 (0.66)	20	21.67 (3.95)	20.27 (4.84)	20	20.22 (9.54)	20.89 (8.85)
R1	0.01	1	1.0 (0.69)	0.78 (1.15)	20	1.09 (1.55)	19.74 (4.75)	20	1.0 (0.8)	18.6 (26.71)
R1n1000	0.01	1	1.0 (0.67)	0.76 (1.13)	20	1.0 (1.36)	19.8 (2.11)	20	1.16 (0.89)	21.33 (22.36)
R2	0.1	1	1.0 (0.37)	1.0 (0.89)	50	51.0 (8.9)	51.25 (18.16)	50	51.0 (17.05)	51.0 (42.4)
R2	0.01	1	1.59 (1.62)	1.0 (0.74)	50	8.0 (10.41)	50.0 (2.97)	50	2.0 (2.85)	51.0 (26.19)
R3	0.1	1	1.0 (0.74)	1.6 (1.85)	80	90.0 (29.65)	82.0 (34.1)	80	83.67 (40.46)	60.33 (61.28)
R3	0.01	1	1.6 (2.19)	1.0 (0.74)	80	2.5 (3.35)	83.0 (10.38)	80	6.11 (8.73)	80.0 (34.59)
AFR1	0.2	1	2.2 (2.92)	0.75 (0.98)	5	9.8 (13.05)	4.88 (6.85)	5	4.13 (5.11)	5.33 (2.69)
AFR1n1000	0.2	1	2.22 (2.97)	0.67 (0.83)	5	10.44 (14.0)	4.5 (6.33)	5	3.55 (4.33)	5.0 (2.58)
AFR2	0.25	16.67	10.0 (5.93)	6.0 (2.97)	166.67	65.67 (36.08)	33.33 (13.1)	10	8.41 (2.05)	5.6 (0.89)
D1	0.1	20	16.8 (8.43)	15.37 (9.18)	20	20.0 (15.12)	17.45 (17.58)	1	1.14 (1.14)	0.98 (0.79)
D1	0.01	20	47.0 (43.88)	22.5 (5.19)	20	50.0 (64.2)	22.22 (28.66)	1	1.14 (1.24)	1.01 (0.44)
D1n1000	0.01	20	50.0 (47.15)	20.67 (2.35)	20	62.71 (73.82)	20.0 (14.66)	1	1.11 (1.2)	1.01 (0.19)
D2	0.1	50	31.0 (29.65)	22.22 (19.01)	50	30.83 (35.63)	22.22 (31.46)	1	1.09 (1.25)	0.86 (0.74)
D2	0.01	50	95.0 (66.72)	59.0 (21.87)	50	113.13 (167.72)	62.75 (92.22)	1	1.96 (1.42)	1.1 (1.58)
D3	0.1	80	50.0 (49.42)	27.71 (27.15)	80	49.0 (66.42)	22.88 (33.47)	1	1.08 (1.31)	0.84 (0.71)
D3	0.01	80	95.0 (44.48)	84.0 (38.55)	80	113.75 (168.65)	70.0 (103.78)	1	1.91 (1.85)	1.01 (1.49)
AFD1	0.05	5	5.0 (4.45)	5.5 (1.73)	5	10.0 (10.04)	10.0 (5.93)	1	1.88 (0.88)	1.68 (1.01)
AFD1n1000	0.05	5	5.0 (2.8)	5.33 (0.99)	5	9.0 (5.07)	9.33 (3.95)	1	1.82 (0.74)	1.55 (0.84)
AFD2	0.25	166.67	8.83 (5.07)	4.75 (2.8)	166.67	14.92 (18.16)	3.0 (4.45)	1	1.9 (2.0)	1.17 (1.68)
A1	0.1	10	10.2 (1.19)	10.5 (4.98)	20	18.0 (5.19)	19.53 (10.3)	2	1.88 (0.8)	1.84 (0.88)
A1	0.01	10	9.4 (3.11)	12.35 (4.97)	20	17.0 (11.86)	22.22 (29.32)	2	1.88 (0.76)	1.72 (0.75)
A1n1000	0.01	10	10.0 (1.98)	10.8 (1.78)	20	18.2 (13.39)	20.0 (25.87)	2	1.89 (1.32)	1.77 (0.55)
A2	0.1	20	18.8 (10.91)	10.47 (4.16)	50	20.0 (19.71)	22.22 (19.45)	2.5	1.11 (1.14)	2.13 (0.95)
A2	0.01	20	49.0 (46.26)	21.67 (4.02)	50	56.0 (71.17)	22.22 (13.19)	2.5	1.19 (1.17)	1.03 (0.19)
A3	0.1	50	39.67 (33.31)	18.8 (15.12)	80	47.0 (56.19)	24.19 (32.3)	1.6	1.12 (1.24)	1.41 (0.8)
A3	0.01	50	92.75 (61.9)	55.0 (18.29)	80	111.25 (164.94)	81.0 (97.11)	1.6	1.96 (1.43)	1.44 (1.69)
AFA1	0.1	4.33	5.0 (3.9)	4.89 (1.91)	7.67	10.0 (9.14)	8.89 (5.35)	1.77	1.89 (0.73)	1.83 (0.99)
AFA1n1000	0.1	4.33	5.0 (2.67)	4.5 (1.07)	7.67	9.0 (3.56)	8.4 (3.56)	1.77	1.8 (0.36)	1.89 (0.87)
AFA2	0.5	83.33	5.0 (4.94)	8.58 (12.19)	166.67	10.0 (10.91)	24.54 (36.38)	2	1.85 (0.67)	1.78 (1.16)
U1	0.1	20	16.0 (10.5)	25.0 (17.3)	1	15.0 (21.99)	1.19 (1.66)	0.05	1.05 (1.56)	0.07 (0.1)
U1	0.01	20	46.0 (43.0)	23.5 (6.67)	1	49.0 (61.68)	22.22 (30.87)	0.05	1.14 (1.24)	0.75 (0.51)
U2	0.1	50	40.0 (50.41)	118.75 (146.41)	1	29.33 (39.78)	1.0 (1.48)	0.02	0.59 (0.87)	0.01 (0.01)
U2	0.01	50	99.17 (72.15)	71.67 (33.36)	1	113.75 (168.65)	17.75 (26.32)	0.02	1.96 (2.09)	0.27 (0.4)
U3	0.1	80	80.0 (107.49)	272.5 (375.99)	1	33.33 (47.94)	1.0 (1.48)	0.0125	0.38 (0.57)	0.0 (0.0)
U3	0.01	80	98.0 (47.07)	112.5 (67.95)	1	116.25 (172.35)	11.13 (16.49)	0.0125	1.92 (2.85)	0.16 (0.24)
U3n1000	0.01	80	90.0 (33.36)	100.0 (37.07)	1	100.0 (148.26)	2.75 (4.08)	0.0125	1.16 (1.72)	0.03 (0.04)
U4	0.35	90	1.67 (2.4)	0.19 (0.28)	1	16.08 (23.08)	1.2 (1.16)	0.0111	13.33 (19.64)	0.05 (11.78)

Ratios with both the numerator and denominator being exactly 0 were set to 1, ratios with only the denominator being exactly 0 were set to the arbitrarily chosen high number 10^6 to include their information for the calculation of the median.

MAD, median absolute deviation, adjusted by a constant (1.4826) for asymptotically normal consistency; P, true value for the disease allele frequency; ASP, affected sib pair; AST, affected sib triplet; f_i/f_0, estimated penetrance ratio; F_i/F_0, true value for penetrance ratio.

Table 5. Results of the imprinting models

Model name	P	F		GHM analysis with imprinting							GHM analysis without imprinting				
				Estimated median parameters (MAD)							Estimated median parameters (MAD)				
				MOD	p	f_0	$f_{1,pat}$	$f_{1,mat}$	f_2	i	MOD	p	f_0	f_1	f_2
cpi ($I = -1$)	0.01	0; 0; 1; 1	ped 1	26.0 (2.66)	0.008 (0.003)	0.0002 (0.0003)	0.0003 (0.0004)	0.94 (0.089)	0.94 (0.089)	–1 (0.0)	9.26 (2.42)	0.1 (0.0)	0.01 (0.0)	0.54 (0.615)	0.5 (0.638)
			ped 2	39.61 (2.81)	0.003 (0.004)	0.0 (0.0)	0.0 (0.0)	0.95 (0.074)	0.555 (0.645)	–0.1 (1.28)	39.49 (2.93)	0.002 (0.0)	0.0001 (0.0001)	0.43 (0.163)	1.0 (0.0)
cmi ($I = 1$)	0.01	0; 1; 0; 1	ped 3	24.17 (2.63)	0.008 (0.003)	0.0003 (0.0004)	0.9 (0.148)	0.001 (0.002)	0.7 (0.445)	0.93 (0.13)	18.81 (2.78)	0.05 (0.01)	0.0003 (0.0004)	0.12 (0.059)	0.51 (0.415)
ni	0.01	0; 0.5; 0.5; 1	ped 1	43.89 (4.0)	0.01 (0.0)	0.0002 (0.0003)	0.71 (0.423)	0.76 (0.356)	0.94 (0.089)	0 (0.44)	43.86 (3.99)	0.01 (0.0)	0.0 (0.0)	0.2 (0.274)	0.9 (0.148)
			ped 2	34.45 (3.48)	0.008 (0.003)	0.0005 (0.0007)	0.38 (0.561)	0.43 (0.415)	0.92 (0.119)	0 (0.76)	34.35 (3.4)	0.008 (0.0)	0.0005 (0.0004)	0.48 (0.356)	0.83 (0.252)
			ped 3	27.06 (3.17)	0.008 (0.003)	0.0004 (0.0006)	0.001 (0.001)	0.44 (0.474)	0.67 (0.489)	–0.035 (0.85)	26.89 (3.12)	0.008 (0.01)	0.0003 (0.0004)	0.48 (0.334)	0.85 (0.222)

cpi, complete paternal imprinting; cmi, complete maternal imprinting; ni, no imprinting; MAD, median absolute deviation, adjusted by a constant (1.4826) for asymptotically normal consistency; *I*, true value for the imprinting index; *P*, true value for the disease allele frequency; *F*, true values for the penetrances (F_0; $F_{1,pat}$; $F_{1,mat}$; F_2); ped, pedigree structure; MOD, MOD score; *p*, estimated disease allele frequency; f_i, estimated penetrances; *i*, estimated imprinting index.

models were mostly estimated with good accuracy, whereas DST and DSQ data sometimes showed larger bias than MAD for specific parameters.

(2) As was expected, the estimation of trait-model parameters and penetrance ratios improved when adding an affected sibling to an ASP, resulting in an AST. The identifiability of the trait-model type depended on the true point of allele-sharing in the corresponding parameter space. The parameter space for ASPs is the possible triangle, whereas the parameter space for ASTs has not been graphically depicted so far. However, using the formulas given by Knapp [28], we were able to empirically draw the parameter space for ASTs (Fig. 1), and hence to hypothesize which model types could be better discriminated using ASTs compared to ASPs. As was expected from the structure of the parameter spaces for both pedigree types, estimation accuracy using ASTs was particularly higher for overdominant models compared to ASPs. Discrimination of additive and dominant models, especially when the genetic effect was small to moderate, remained difficult. Recessive models were generally identified as such using either ASPs or ASTs due to their clear spatial separation in the parameter space from other model types.

(3) In line with our expectations, the identifiability of absolute values of the penetrances instead of pairwise ratios could be achieved when unaffected pedigree members were included in the analysis, i.e., DSTs and DSQs as well as 3-G pedigrees.

(4) Interestingly, the identifiability of trait-model parameters was only slightly better when adding a further unaffected sibling to DSTs, i.e., when using DSQs. The number of allele-sharing classes of DSTs hence seemed to be sufficient for the identification of the trait-model parameters.

(5) With more complex pedigrees, the identifiability of trait-model parameters further improved for some models. While the median estimates were mostly similar, using 3-G pedigrees instead of DSTs or DSQs often led to a reduction in MAD of the parameters.

Imprinting Models

The results of the imprinting scenarios can be found in Table 5. All parental genotypes were removed for both AHSPs and ASPs prior to the analysis.

NI Model

Using pedigree structure 1, i.e., AHSPs with one half of the sample having a common father and the other half having a common mother, the disease allele fre-

quency P and the penetrance F_0 were estimated with high accuracy for the ni model in a MOD score analysis without taking imprinting into account. However, penetrances F_1 and F_2 were both underestimated, with more downward bias for F_1. It is of note that only 1 free parameter can in principle be identified from AHSP data in a MOD score analysis. In the case of the corresponding analysis taking imprinting into account, P and F_0 were estimated with high accuracy. Penetrance F_2 was estimated close to its expected value; however, the heterozygote penetrances were both clearly overestimated. The median values for the heterozygote penetrances $F_{1,pat}$, and $F_{1,mat}$ were comparable, which was expected for the ni model. The correct imprinting index $I = 0$ was obtained in the analysis of pedigree structure 1 and the ni model. MOD scores were comparable between the 2 analyses, i.e., with and without taking imprinting into account, whereby the imprinting MOD score is per definition always as large as the corresponding ni score. In the case of the ni model, MOD scores were generally highest using pedigree structure 1 and lowest for pedigree structure 3.

Using pedigree structure 2, i.e., 100 ASPs and 100 AHSPs having a common mother, the estimated median disease allele frequency P and penetrances F_0 and F_1 were close to the expected value in the analysis without taking imprinting into account. Penetrance F_2 was underestimated. In the case of the analysis taking imprinting into account, P and F_0 were estimated close to the expected value, whereas the heterozygote penetrances $F_{1,pat}$ and $F_{1,mat}$ as well as F_2 were underestimated. As was with pedigree structure 1, the correct imprinting index $I = 0$ could be obtained from the analysis of pedigree structure 2. MOD scores of both analysis types were comparable.

The corresponding values for the trait-model parameters for pedigree structure 3, i.e., 180 ASPs and 20 AHSPs with a common mother, were comparable to those of pedigree structure 2 for the ni analysis. When imprinting was taken into account in the analysis, penetrances $F_{1,pat}$ and F_2 were estimated lower ($f_{1,pat} = 0.001$; $f_2 = 0.67$) compared to pedigree structure 2 ($f_{1,pat} = 0.38$; $f_2 = 0.92$). Most strikingly, penetrance $F_{1,pat}$ was estimated close to 0, which reflects the unidentifiability between paternal imprinting and ni models when parental genotypes have been removed. It appears counterintuitive at first sight that an apparently stronger indication of paternal imprinting is obtained for pedigree structure 3, which contains only 20 AHSPs, compared to pedigree structure 2, which contains 100 AHSPs (Table 5). However, with a larger number of AHSPs in pedigree structure 2, it is more likely that 2 half-sibs have received the disease allele from the 2 separate fathers rather than from their joint mother, which reduces the likelihood of a paternal imprinting model. The estimated median imprinting index was estimated close to its expected value, albeit slightly below 0 due to the underestimation of $F_{1,pat}$.

Imprinting Models

In contrast to the ni model, the presentation of the results for the imprinting simulations starts with the MOD scores taking imprinting into account, which are then compared to the ni results. Using pedigree structure 1 and the cpi model, the disease allele frequency and the penetrances were estimated with good accuracy in a MOD score analysis taking imprinting into account. The correct imprinting index $I = -1$ could be obtained as well. With regard to the corresponding ni analysis, the median estimated trait-model parameters of the cpi model were difficult to interpret due to the following: since the ni MOD score analysis assumes the equivalence of parental genomes, i.e., the equivalence of AHSPs having a common father and AHSPs having a common mother, this leads to a reduced likelihood and to bias of trait-model parameter estimates. This is because the truly underlying genetic mechanism, i.e., the imparity of parental genomes, is misspecified in a ni MOD score analysis, which cannot be compensated by maximizing over the ni trait model. If complete imprinting is really present but not modelled in the analysis, only the meioses of those AHSPs with a common parent of the non-imprinted sex contribute linkage information, whereas the other AHSPs point at no linkage. Therefore, the MOD score dropped from 26.0 with imprinting to 9.26 without imprinting taken into account in the analysis, and trait-model parameter estimates for the ni model were distorted.

Using pedigree structure 2, trait-model parameters could be estimated with good accuracy in an imprinting MOD score analysis, except for F_2, which was clearly underestimated. In fact, F_2 was mostly estimated as either 0 or close to 1 (data not shown). This was most likely due to the fact that a paternal imprinting model with penetrances (F_0, $F_{1,pat}$, $F_{1,mat}$, F_2) = (0;0;1;1) can hardly be distinguished from an overdominant model with penetrances (0;0;1;0) using ASP data. This was also reflected by a median imprinting index with a smaller absolute value than expected ($i = -0.1$), because i is defined to be 0 if the estimates of F_0 and F_2 are equal, and a high MAD for the F_2 penetrance (0.645; Table 5). Owing to the AHSPs with a common mother, however, the relation $F_{1,pat} << F_{1,mat}$ could mostly be determined. With regard to the corre-

sponding ni analysis, trait-model parameters were estimated with good accuracy, whereby the median heterozygote penetrance f_1 was estimated close to the mean of the penetrances $F_{1,pat}$ and $F_{1,mat}$ that were used for the cpi model simulation. MOD scores of the ni analysis were comparable to those of the imprinting analysis for pedigree structure 2, because assuming strong maternal allele sharing is almost as likely as an additive model, for which allele sharing can take place through parents of both sexes. In other words, maternal allele sharing in AHSPs with a common mother does not imply random (non-excess) paternal allele sharing in ASPs with untyped parents.

Using pedigree structure 3 and the cmi model, the combined sample of 180 ASPs and 20 AHSPs having a common mother led to trait-model parameter estimates reflecting maternal imprinting, albeit with an underestimation of $F_{1,pat}$ and F_2 ($f_{1,pat} = 0.9$; $f_2 = 0.7$). The reason why F_2 was underestimated is the same as it was for pedigree structure 2. In contrast to pedigree structure 2, the imprinting analysis yielded substantially higher MOD scores than the ni analysis, because the non-excess allele sharing of AHSPs with a common mother can only be explained by maternal imprinting, whereas for the ni analysis the non-excess sharing of maternal alleles in AHSPs reduces linkage information. This goes along with distorted trait-model parameter estimates for the combined dataset. The imprinting index I was estimated close to its expected value reflecting maternal imprinting.

Summary of Imprinting Results

The imprinting results are summarized as an answer to question (6) given in the Introduction section.

Imprinting could reliably be detected in samples that include AHSPs having a common father as well as AHSPs with a common mother, even if the parents are untyped (pedigree structure 1). When analyzing an equal mixture of ASPs and AHSPs having a common mother, all with untyped parents, imprinting could in part be declared when looking at the imprinting index I obtained from the imprinting MOD score analysis and the cpi model. However, the difference between the ni and imprinting MOD score seemed to be marginal, such that there was no significant evidence for imprinting. However, using 180 ASPs and 20 AHSPs having a common mother, again with untyped parents, the results for the cmi model clearly showed that information on imprinting can be extracted when adding a few AHSPs with a common parent of the imprinted sex to a sample of ASPs with untyped parents, which only harbor information on linkage, to obtain substantial evidence of imprinting.

Discussion

The ability of a pedigree analysis to estimate parameters of trait inheritance has been extensively discussed in the literature [1, 11, 16, 19–23]. More specifically, the possibility to jointly estimate linkage and segregation parameters in a MOD score analysis has been debated. A MOD score analysis does not perform classical segregation analysis in the sense of determining whether or not there is major gene segregation, but it estimates some segregation-model parameters together with parameters for linkage, which we denote joint trait-marker inheritance parameters (recombination fraction, LD parameters, and trait-model parameters: disease allele frequency and penetrances). Since the publication of the AAF method proposed by Ewens and Shute [17], the MOD score has often been referred to as being AAF, such that it delivers asymptotically unbiased estimates of the trait-model parameters [11]. It is of note that estimates obtained from maximum likelihood techniques are naturally biased for finite sample sizes. However, the problem of ascertainment or sampling was often neglected and most theoretical work on parameter estimation in linkage analysis assumed what is called PI sampling, i.e., sampling of fixed pedigree structures independent of any proband. Hence, if no correction of the likelihood as to the ascertainment procedure is applied, the estimates of the joint marker-trait inheritance parameters will be biased.

Over the years, the problem of ascertainment/sampling for linkage analysis was gradually elaborated [1, 16, 19–23]. Presumably the most comprehensive and most detailed work on these aspects of pedigree analysis is the book by Ginsburg et al. [1], who claimed that unbiased estimates can in fact be obtained from a pedigree analysis (see also Ginsburg et al. [16]). They provided a general likelihood framework that can be used to accommodate the likelihood for many aspects of the sampling procedure and also showed how to accomplish sampling correction in practice. Although their focus was not on the MOD score approach per se, they provided the abovementioned conditions (i)–(vii), under which the MOD score delivers asymptotically unbiased parameter estimates. Along these lines, one of the goals of the present paper was to investigate the ability of a MOD score analysis to obtain unbiased trait-model parameter estimates in practical situations. To this end, we have thoroughly recapitulated the theoretical background, including conditions under which the parameter estimates should be asymptotically unbiased. We then evaluated the parameter estimation performance of a MOD score analysis in

a simulation study. The first condition of correctly specifying the mode of inheritance referring to the number of loci and the number of alleles at each locus is presumably most crucial. Therefore, a diallelic autosomal binary trait locus was used for the simulation of pedigree data, which is usually assumed as the mode of inheritance in a MOD score analysis. Although complex disorders are expected to follow more complicated modes of inheritance, e.g., involving a larger number of trait loci, the number of possible models, i.e., degrees of freedom, to be tested in a MOD score analysis would be prohibitively large and procedures to avoid inflated type I error rates would presumably diminish power. The second condition of marker-independent sampling is often ignored in practice. However, when performing linkage analysis in the era of densely available markers, this assumption is likely to hold, since the limiting step is often the recruitment of an individual rather than obtaining informative genotypes. Conditions (iii)–(vi) refer to the sampling procedure, which is often assumed to be PI. Admittedly, only few linkage studies are really PI. However, even in this case, parameter estimation remains free of bias if the sampling procedure can be controlled. This is the case if either all members of the PSF have measured trait values (see condition [v]), e.g., by using a questionnaire to include information on potential probands not sampled (see Ginsburg et al. [16]), or sampling is single in the sense of Hodge and Vieland [20] (see condition [vi]), and the model of extension is random (see condition [iv]). Then, the MOD score can readily be used to obtain asymptotically unbiased joint marker-trait inheritance parameter estimates.

Previous work using simulated pedigree data has shown that the maximum LOD score is obtained for the truly underlying genetic model, provided that there is enough power to detect linkage [43]. However, the focus of the aforementioned work was only on strictly dominant ($f_1 = f_2$) and strictly recessive models ($f_1 = 0$) without phenocopies ($f_0 = 0$) and with the disease allele frequency fixed at the true value for the analysis. In addition, maximization was done using a limited set of penetrance values [43]. In our simulation study, the MOD score with a more exhaustive maximization as implemented in GHM was used. Furthermore, we studied a wider range of trait models and pedigree structures. We did not investigate the ability of the MOD score to estimate the recombination fraction and any LD parameters. The recombination fraction is confounded with the trait-model parameters, i.e., with the disease allele frequency p and the 3 penetrances f_0, f_1, and f_2, and was hence excluded from the estimation, but rather fixed at the true value of $\theta = 0$. Otherwise, it would not be possible to distinguish confounding of parameters and bias from each other. In the current program version of GHM, LD is not modelled. As stated earlier, to obtain unbiased trait-model parameter estimates, LD between markers and disease locus must in fact be absent, otherwise sampling is no longer marker independent. As noted by Malkin and Elston [19], such a situation is unlikely when using marker panels of densely spaced single nucleotide polymorphisms. However, selective inclusion of only a subset of markers can ensure linkage equilibrium at least between these markers, while still retaining sufficient information for linkage analysis. With such a sparser set of markers, it is also less likely that one of them is in LD with a disease allele. If LD between marker and disease alleles happens to be present, the expected bias in parameter estimates is so far unknown. Further, we did not consider bias of trait-model parameters due to gene-environment interactions, which are usually not controlled in a linkage analysis. In addition, we did not investigate the ascertainment or sampling bias that may occur when recruiting families in practice. Still, the problem of ascertainment or sampling for linkage analysis with estimation of joint trait-marker inheritance parameters has been thoroughly reviewed and discussed in the Introduction section.

Another aspect of estimating trait-model parameters is their identifiability. It has been shown by Strauch [10] that the identifiability of trait-model parameters depends on the truly underlying number of allele-sharing classes. In addition, only penetrance ratios can be estimated from affecteds-only data. The identifiability is expected to increase with larger sibships or more complex pedigrees. Therefore, we were interested in the degree to which the identifiability of trait-model parameters increases when adding affected or unaffected siblings to an ASP or when analyzing a 3-G pedigree.

In this study, we were able to show how trait-model parameters can in principle be estimated in a MOD score linkage analysis and to what extent the identifiability depends on the pedigree types in the dataset. Our findings can provide guidance to researchers aiming to estimate parameters by a MOD score linkage analysis using family data. Parameter estimation generally showed smaller bias and MAD with increasing pedigree complexity for all investigated model types. Identifiability of trait-model parameters increased with (a) more affected siblings in an affecteds-only analysis of nuclear families, although only ratios of parameter values can be identified in this case, (b) adding unaffected siblings to nuclear families, and for some models with (c) adding a generation (3-G pedi-

grees). Penetrance estimation performance was substantially affected by confounding of the trait-model parameters in terms of their proximity or identity in the corresponding nonparametric allele-sharing parameter space. This is equivalent to the more "parametric" notion that the degree of information to accurately estimate parameters given their identifiability still depends on the proportions of disease locus genotypes that are induced by the number of affected and unaffected individuals in a pedigree, together with the truly underlying trait-model parameters. Therefore, especially additive and dominant models can hardly be distinguished, even when analyzing more complex pedigrees. A sufficient number of pedigrees in the sample is a further prerequisite to be able to actually estimate the parameters in practice, according to the identifiability that is theoretically possible with a certain pedigree type. Furthermore, we have shown under which scenarios imprinting can be detected even if all parents have missing genotypes. Imprinting could reliably be estimated in terms of the imprinting index *I* [35] with the datasets containing both AHSPs having a common father as well as a common mother. We were also able to show that it is possible to combine pure linkage information from ASPs with imprinting-sensitive linkage information from AHSPs having a common mother to obtain substantial evidence for maternal imprinting. This finding indicates that adding AHSPs with a common parent of the imprinted sex draws the trait-model parameter estimates of the combined ASP/AHSP sample towards the truly underlying imprinting model.

In essence, asymptotically unbiased parameter estimates can be obtained from a MOD score analysis, given that certain conditions are satisfied ([i]–[vii], see Introduction section). In most real-life situations, these conditions can hardly be fulfilled. The extent to which a violation of any of these conditions or a combination of them causes bias is unclear and demands further investigations. Such a subsequent simulation study might reveal situations in which, despite, for example, an incorrect sampling model, the parameter estimates obtained from the analysis are essentially correct, which has been referred to as the "man bites dog" criterion [11]. Along these lines, the results of our present study are an important prerequisite for future investigations on robustness of MOD score-based parameter estimation under various sampling schemes.

Acknowledgements

This work was supported by grants Str643/4-1 and Str643/6-1 of the Deutsche Forschungsgemeinschaft (German Research Foundation). Further, this research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. In addition, we greatly appreciate the reviewers' thoughtful comments, which have helped to improve the paper.

Appendix

Table A1. Estimation of trait-model parameters for ASPs and ASTs (recessive models)

Model name	P	Estimated median p (MAD)		F_0	Estimated median f_0 (MAD)		F_1	Estimated median f_1 (MAD)		F_2	Estimated median f_2 (MAD)		D	Estimated median d (MAD)	
		ASP	AST		ASP	AST		ASP	AST		ASP	AST		ASP	AST
R1	0.1	0.1 (0.01)	0.1 (0.03)	0.01	0.04 (0.007)	0.045 (0.007)	0.01	0.045 (0.022)	0.045 (0.022)	0.2	0.92 (0.074)	0.92 (0.07)	–1.0	–1.0 (0.06)	–1.0 (0.07)
R1	0.01	0.015 (0.02)	0.01 (0.0)	0.01	0.045 (0.052)	0.045 (0.007)	0.01	0.05 (0.059)	0.015 (0.022)	0.2	0.13 (0.178)	0.93 (0.074)	–1.0	0.0 (0.0)	–1.05 (0.12)
R1n1000	0.01	0.05 (0.06)	0.01 (0.0)	0.01	0.05 (0.059)	0.045 (0.007)	0.01	0.08 (0.104)	0.02 (0.03)	0.2	0.14 (0.193)	0.94 (0.044)	–1.0	0.0 (0.06)	–1.03 (0.12)
R2	0.1	0.1 (0.01)	0.1 (0.01)	0.01	0.01 (0.0)	0.01 (0.003)	0.01	0.01 (0.003)	0.01 (0.007)	0.5	0.5 (0.044)	0.51 (0.089)	–1.0	–1.0 (0.01)	–1.0 (0.04)
R2	0.01	0.1 (0.14)	0.01 (0.0)	0.01	0.045 (0.052)	0.01 (0.0)	0.01	0.1 (0.136)	0.01 (0.007)	0.5	0.48 (0.341)	0.5 (0.03)	–1.0	–0.81 (1.2)	–1.0 (0.03)
R3	0.1	0.1 (0.01)	0.1 (0.0)	0.01	0.01 (0.003)	0.008 (0.003)	0.01	0.01 (0.007)	0.01 (0.013)	0.8	0.89 (0.111)	0.74 (0.326)	–1.0	–1.0 (0.01)	–0.99 (0.04)
R3	0.01	0.48 (0.09)	0.01 (0.0)	0.01	0.05 (0.059)	0.01 (0.0)	0.01	0.1 (0.078)	0.01 (0.007)	0.8	0.5 (0.593)	0.85 (0.089)	–1.0	–0.64 (0.95)	–1.0 (0.01)
AFR1	0.2	0.44 (0.22)	0.17 (0.18)	0.04	0.05 (0.059)	0.07 (0.067)	0.04	0.1 (0.059)	0.09 (0.044)	0.2	0.49 (0.474)	0.49 (0.104)	–1.0	–0.56 (0.82)	–0.80 (0.56)
AFR1n1000	0.2	0.45 (0.12)	0.21 (0.23)	0.04	0.05 (0.059)	0.08 (0.089)	0.04	0.11 (0.059)	0.09 (0.03)	0.2	0.5 (0.193)	0.48 (0.163)	–1.0	–0.55 (2.02)	–0.77 (0.64)
AFR2	0.25	0.07 (0.04)	0.1 (0.06)	0.003	0.01 (0.003)	0.015 (0.007)	0.05	0.09 (0.044)	0.1 (0.015)	0.5	0.57 (0.489)	0.52 (0.044)	–0.81	–0.79 (0.06)	–0.69 (0.07)

MAD, median absolute deviation, adjusted by a constant (1.4826) for asymptotically normal consistency; p, estimated value for the disease allele frequency; f_0, f_1, f_2, estimated values for penetrances with f_i denoting the probability that an individual with i copies of the disease allele is affected by the disease; d, estimated value for the dominance index; P, true value for the disease allele frequency; F_i, true values for the penetrances; D, true value for the dominance index; ASP, affected sib pair; AST, affected sib triplet.

Table A2. Estimation of trait-model parameters for ASPs and ASTs (dominant models)

Model name	P	Estimated median p (MAD)		F_0	Estimated median f_0 (MAD)		F_1	Estimated median f_1 (MAD)		F_2	Estimated median f_2 (MAD)		D	Estimated median d (MAD)	
		ASP	AST		ASP	AST		ASP	AST		ASP	AST		ASP	AST
D1	0.1	0.1 (0.06)	0.1 (0.01)	0.01	0.05 (0.007)	0.045 (0.022)	0.2	0.63 (0.371)	0.555 (0.363)	0.2	0.74 (0.385)	0.655 (0.511)	1.0	0.39 (1.38)	0.85 (1.39)
D1	0.01	0.1 (0.0)	0.01 (0.01)	0.01	0.01 (0.014)	0.04 (0.015)	0.2	0.52 (0.415)	0.92 (0.074)	0.2	0.86 (0.208)	0.925 (0.111)	1.0	0.11 (1.03)	0.91 (1.01)
D1n1000	0.01	0.1 (0.0)	0.008 (0.01)	0.01	0.01 (0.013)	0.045 (0.007)	0.2	0.53 (0.445)	0.93 (0.044)	0.2	0.88 (0.178)	0.95 (0.074)	1.0	0.66 (0.98)	0.90 (0.39)
D2	0.1	0.1 (0.01)	0.1 (0.01)	0.01	0.01 (0.014)	0.015 (0.022)	0.5	0.565 (0.482)	0.61 (0.356)	0.5	0.57 (0.571)	0.53 (0.549)	1.0	0.66 (1.17)	0.96 (1.45)
D2	0.01	0.025 (0.03)	0.01 (0.0)	0.01	0.005 (0.006)	0.008 (0.003)	0.5	0.48 (0.119)	0.47 (0.133)	0.5	0.52 (0.638)	0.18 (0.267)	1.0	0.0 (0.86)	–0.04 (1.19)
D3	0.1	0.1 (0.0)	0.1 (0.01)	0.01	0.01 (0.013)	0.01 (0.015)	0.8	0.53 (0.474)	0.6 (0.371)	0.8	0.55 (0.593)	0.52 (0.578)	1.0	0.51 (1.28)	0.96 (1.44)
D3	0.01	0.01 (0.0)	0.01 (0.0)	0.01	0.004 (0.005)	0.006 (0.006)	0.8	0.49 (0.563)	0.49 (0.311)	0.8	0.50 (0.652)	0.26 (0.385)	1.0	0.01 (0.98)	–0.03 (1.37)
AFD1	0.05	0.1 (0.13)	0.11 (0.03)	0.04	0.025 (0.03)	0.015 (0.013)	0.2	0.1 (0.074)	0.12 (0.104)	0.2	0.33 (0.356)	0.18 (0.193)	1.0	0.0 (0.49)	0.0 (0.64)
AFD1n1000	0.05	0.1 (0.07)	0.11 (0.03)	0.04	0.04 (0.044)	0.015 (0.007)	0.2	0.1 (0.074)	0.1 (0.074)	0.2	0.23 (0.252)	0.155 (0.111)	1.0	0.0 (0.39)	0.18 (0.69)
AFD2	0.25	0.1 (0.13)	0.15 (0.1)	0.003	0.05 (0.059)	0.02 (0.03)	0.5	0.1 (0.059)	0.1 (0.104)	0.5	0.45 (0.534)	0.12 (0.177)	1.0	0.0 (1.09)	0.11 (2.25)

For more details, see Appendix Table A1.

Table A3. Estimation of trait-model parameters for ASPs and ASTs (additive models)

Model name	P	Estimated median p (MAD)		F_0	Estimated median f_0 (MAD)		F_1	Estimated median f_1 (MAD)		F_2	Estimated median f_2 (MAD)		D	Estimated median d (MAD)	
		ASP	AST		ASP	AST		ASP	AST		ASP	AST		ASP	AST
A1	0.1	0.1 (0.04)	0.1 (0.01)	0.01	0.05 (0.007)	0.04 (0.03)	0.1	0.49 (0.044)	0.42 (0.163)	0.2	0.875 (0.185)	0.91 (0.129)	−0.05	0.0 (0.67)	0.02 (0.47)
A1	0.01	0.05 (0.07)	0.05 (0.06)	0.01	0.045 (0.052)	0.025 (0.034)	0.1	0.11 (0.089)	0.49 (0.133)	0.2	0.47 (0.563)	0.93 (0.104)	−0.05	0.0 (0.62)	0.03 (0.47)
A1n1000	0.01	0.01 (0.01)	0.03 (0.03)	0.01	0.01 (0.015)	0.015 (0.022)	0.1	0.1 (0.03)	0.49 (0.074)	0.2	0.45 (0.519)	0.94 (0.089)	−0.05	0.0 (0.67)	−0.01 (0.33)
A2	0.1	0.09 (0.04)	0.07 (0.04)	0.01	0.045 (0.007)	0.04 (0.03)	0.2	0.79 (0.297)	0.45 (0.119)	0.5	0.87 (0.193)	0.95 (0.074)	−0.22	0.66 (0.99)	−0.12 (0.31)
A2	0.01	0.1 (0.0)	0.008 (0.01)	0.01	0.01 (0.013)	0.045 (0.007)	0.2	0.50 (0.267)	0.93 (0.059)	0.5	0.89 (0.163)	0.96 (0.059)	−0.22	0.01 (0.99)	0.89 (0.39)
A3	0.1	0.1 (0.01)	0.1 (0.01)	0.01	0.01 (0.014)	0.035 (0.037)	0.5	0.55 (0.43)	0.50 (0.133)	0.8	0.86 (0.208)	0.87 (0.193)	0.24	0.54 (0.82)	0.30 (0.67)
A3	0.01	0.02 (0.02)	0.01 (0.0)	0.01	0.005 (0.006)	0.008 (0.003)	0.5	0.48 (0.133)	0.44 (0.178)	0.8	0.52 (0.638)	0.19 (0.282)	0.24	0.0 (0.85)	−0.07 (0.8)
AFA1	0.1	0.1 (0.09)	0.11 (0.09)	0.03	0.023 (0.033)	0.02 (0.03)	0.13	0.11 (0.089)	0.16 (0.089)	0.23	0.325 (0.348)	0.32 (0.282)	0.0	0.0 (0.38)	−0.08 (0.55)
AFA1n1000	0.1	0.1 (0.07)	0.11 (0.06)	0.03	0.015 (0.022)	0.02 (0.03)	0.13	0.1 (0.074)	0.15 (0.074)	0.23	0.235 (0.252)	0.28 (0.267)	0.0	0.0 (0.2)	−0.08 (0.46)
AFA2	0.5	0.14 (0.2)	0.415 (0.14)	0.003	0.04 (0.044)	0.01 (0.015)	0.25	0.15 (0.148)	0.45 (0.423)	0.5	0.46 (0.534)	0.85 (0.219)	<0.01	0.0 (0.1)	0.0 (0.68)

For more details, see Appendix Table A1.

Table A4. Estimation of trait-model parameters for ASPs and ASTs (overdominant models)

Model name	P	Estimated median p (MAD)		F_0	Estimated median f_0 (MAD)		F_1	Estimated median f_1 (MAD)		F_2	Estimated median f_2 (MAD)	
		ASP	AST		ASP	AST		ASP	AST		ASP	AST
U1	0.1	0.1 (0.04)	0.1 (0.01)	0.01	0.035 (0.03)	0.03 (0.03)	0.2	0.5 (0.623)	0.9 (0.147)	0.01	0.36 (0.474)	0.035 (0.052)
U1	0.01	0.1 (0.0)	0.01 (0.01)	0.01	0.01 (0.013)	0.04 (0.015)	0.2	0.51 (0.474)	0.92 (0.074)	0.01	0.825 (0.259)	0.5 (0.704)
U2	0.1	0.1 (0.01)	0.1 (0.01)	0.01	0.01 (0.01)	0.005 (0.007)	0.5	0.47 (0.563)	0.7 (0.385)	0.01	0.41 (0.46)	0.004 (0.006)
U2	0.01	0.025 (0.03)	0.01 (0.0)	0.01	0.004 (0.006)	0.008 (0.003)	0.5	0.48 (0.133)	0.49 (0.133)	0.01	0.50 (0.652)	0.1 (0.148)
U3	0.1	0.1 (0.01)	0.1 (0.01)	0.01	0.006 (0.006)	0.001 (0.001)	0.8	0.49 (0.578)	0.69 (0.385)	0.01	0.175 (0.259)	0.0 (0.0)
U3	0.01	0.01 (0.0)	0.01 (0.0)	0.01	0.003 (0.003)	0.002 (0.003)	0.8	0.47 (0.549)	0.48 (0.43)	0.01	0.50 (0.623)	0.04 (0.059)
U3n1000	0.01	0.01 (0.0)	0.01 (0.0)	0.01	0.006 (0.006)	0.006 (0.006)	0.8	0.58 (0.445)	0.58 (0.385)	0.01	0.44 (0.652)	0.007 (0.01)
U4	0.35	0.1 (0.14)	0.37 (0.1)	0.01	0.01 (0.015)	0.09 (0.133)	0.9	0.05 (0.073)	0.06 (0.089)	0.01	0.39 (0.563)	0.22 (0.326)

For more details, see Appendix Table A1.

Table A5. Estimation of trait-model parameters of DST and DSQ pedigrees (recessive models)

Model name	P	Estimated median p (MAD)		F_0	Estimated median f_0 (MAD)		F_1	Estimated median f_1 (MAD)		F_2	Estimated median f_2 (MAD)		D	Estimated median d (MAD)	
		DST	DSQ		DST	DSQ		DST	DSQ		DST	DSQ		DST	DSQ
R1	0.1	0.05 (0.01)	0.05 (0.04)	0.01	0.01 (0.007)	0.008 (0.01)	0.01	0.01 (0.015)	0.003 (0.004)	0.2	0.32 (0.163)	0.26 (0.222)	−1.0	−0.97 (0.06)	−0.97 (0.06)
R1	0.01	0.043 (0.05)	0.045 (0.05)	0.01	0.043 (0.06)	0.09 (0.133)	0.01	0.05 (0.071)	0.1 (0.148)	0.2	0.07 (0.104)	0.195 (0.289)	−1.0	−0.07 (1.12)	−0.83 (1.15)
R1n1000	0.01	0.05 (0.06)	0.05 (0.06)	0.01	0.04 (0.052)	0.085 (0.123)	0.01	0.05 (0.07)	0.09 (0.13)	0.2	0.07 (0.104)	0.19 (0.282)	−1.0	−0.51 (0.76)	−0.97 (1.05)
R2	0.1	0.09 (0.03)	0.09 (0.03)	0.01	0.01 (0.0)	0.01 (0.0)	0.01	0.01 (0.007)	0.01 (0.007)	0.5	0.51 (0.074)	0.5 (0.059)	−1.0	−1.0 (0.03)	−1.0 (0.03)
R2	0.01	0.08 (0.01)	0.05 (0.06)	0.01	0.025 (0.037)	0.03 (0.043)	0.01	0.05 (0.071)	0.045 (0.066)	0.5	0.44 (0.571)	0.46 (0.593)	−1.0	−0.99 (0.44)	−1.0 (0.35)
R3	0.1	0.09 (0.03)	0.09 (0.03)	0.01	0.01 (0.003)	0.01 (0.003)	0.01	0.01 (0.007)	0.01 (0.007)	0.8	0.81 (0.044)	0.8 (0.03)	−1.0	−1.0 (0.03)	−1.0 (0.03)
R3	0.01	0.06 (0.07)	0.05 (0.06)	0.01	0.04 (0.046)	0.035 (0.04)	0.01	0.04 (0.047)	0.035 (0.047)	0.8	0.8 (0.297)	0.82 (0.267)	−1.0	−1.0 (0.14)	−1.0 (0.12)
AFR1	0.2	0.12 (0.16)	0.11 (0.15)	0.04	0.01 (0.015)	0.01 (0.015)	0.04	0.05 (0.062)	0.04 (0.055)	0.2	0.175 (0.256)	0.15 (0.218)	−1.0	−0.87 (0.76)	−0.92 (0.61)
AFR1n1000	0.2	0.15 (0.21)	0.12 (0.16)	0.04	0.01 (0.015)	0.01 (0.015)	0.04	0.05 (0.059)	0.045 (0.052)	0.2	0.19 (0.27)	0.17 (0.23)	−1.0	−0.87 (0.68)	−0.86 (0.6)
AFR2	0.25	0.15 (0.13)	0.22 (0.18)	0.003	0.006 (0.006)	0.005 (0.006)	0.05	0.06 (0.037)	0.05 (0.03)	0.5	0.53 (0.104)	0.52 (0.074)	−0.81	−0.81 (0.09)	−0.84 (0.12)

For more details, see Appendix Table A1. DST, discordant sib triplet; DSQ, discordant sib quadruplet.

Table A6. Estimation of trait-model parameters of DST and DSQ pedigrees (dominant models)

Model name	P	Estimated median p (MAD)		F_0	Estimated median f_0 (MAD)		F_1	Estimated median f_1 (MAD)		F_2	Estimated median f_2 (MAD)		D	Estimated median d (MAD)	
		DST	DSQ		DST	DSQ		DST	DSQ		DST	DSQ		DST	DSQ
D1	0.1	0.1 (0.06)	0.11 (0.06)	0.01	0.008 (0.004)	0.008 (0.003)	0.2	0.12 (0.089)	0.13 (0.059)	0.2	0.31 (0.252)	0.33 (0.222)	1.0	−0.05 (0.8)	−0.03 (0.86)
D1	0.01	0.1 (0.03)	0.11 (0.03)	0.01	0.002 (0.003)	0.005 (0.006)	0.2	0.14 (0.089)	0.14 (0.074)	0.2	0.31 (0.267)	0.34 (0.208)	1.0	−0.01 (0.53)	−0.03 (0.59)
D1n1000	0.01	0.11 (0.01)	0.11 (0.03)	0.01	0.003 (0.003)	0.005 (0.006)	0.2	0.15 (0.059)	0.15 (0.059)	0.2	0.3 (0.208)	0.34 (0.163)	1.0	<0.01 (0.49)	−0.01 (0.5)
D2	0.1	0.1 (0.01)	0.1 (0.03)	0.01	0.006 (0.007)	0.008 (0.01)	0.5	0.47 (0.089)	0.49 (0.059)	0.5	0.555 (0.378)	0.51 (0.341)	1.0	0.41 (1.12)	0.66 (1.13)
D2	0.01	0.045 (0.01)	0.05 (0.04)	0.01	0.003 (0.004)	0.006 (0.006)	0.5	0.48 (0.074)	0.48 (0.044)	0.5	0.74 (0.385)	0.69 (0.46)	1.0	0.05 (0.41)	0.12 (0.5)
D3	0.1	0.08 (0.04)	0.09 (0.06)	0.01	0.015 (0.016)	0.008 (0.01)	0.8	0.8 (0.044)	0.8 (0.03)	0.8	0.71 (0.43)	0.81 (0.222)	1.0	0.81 (0.91)	0.88 (0.44)
D3	0.01	0.02 (0.02)	0.025 (0.03)	0.01	0.008 (0.004)	0.008 (0.004)	0.8	0.8 (0.03)	0.8 (0.03)	0.8	1.00 (0.0)	0.92 (0.119)	1.0	0.62 (0.15)	0.64 (0.2)
AFD1	0.05	0.08 (0.11)	0.08 (0.11)	0.04	0.01 (0.014)	0.01 (0.014)	0.2	0.1 (0.082)	0.11 (0.096)	0.2	0.23 (0.282)	0.27 (0.334)	1.0	−0.11 (0.84)	−0.17 (0.9)
AFD1n1000	0.05	0.09 (0.12)	0.07 (0.09)	0.04	0.01 (0.014)	0.01 (0.014)	0.2	0.11 (0.089)	0.12 (0.089)	0.2	0.27 (0.297)	0.28 (0.311)	1.0	−0.15 (0.73)	−0.33 (0.7)
AFD2	0.25	0.17 (0.13)	0.16 (0.12)	0.003	0.035 (0.037)	0.03 (0.03)	0.5	0.47 (0.119)	0.49 (0.074)	0.5	0.48 (0.341)	0.41 (0.356)	1.0	0.28 (1.52)	0.67 (1.53)

For more details, see Appendix Tables A1 and A5.

Table A7. Estimation of trait-model parameters of DST and DSQ pedigrees (additive models)

Model name	P	Estimated median p (MAD)		F_0	Estimated median f_0 (MAD)		F_1	Estimated median f_1 (MAD)		F_2	Estimated median f_2 (MAD)		D	Estimated median d (MAD)	
		DST	DSQ		DST	DSQ		DST	DSQ		DST	DSQ		DST	DSQ
A1	0.1	0.06 (0.06)	0.06 (0.07)	0.01	0.01 (0.003)	0.01 (0.003)	0.1	0.1 (0.03)	0.1 (0.03)	0.2	0.17 (0.178)	0.17 (0.17)	–0.05	0.01 (0.56)	0.01 (0.89)
A1	0.01	0.01 (0.01)	0.01 (0.01)	0.01	0.01 (0.003)	0.01 (0.0)	0.1	0.09 (0.059)	0.09 (0.059)	0.2	0.15 (0.193)	0.17 (0.222)	–0.05	–0.12 (0.81)	–0.35 (0.7)
A1n1000	0.01	0.01 (0.01)	0.01 (0.01)	0.01	0.01 (0.0)	0.01 (0.0)	0.1	0.1 (0.03)	0.1 (0.044)	0.2	0.165 (0.23)	0.17 (0.23)	–0.05	–0.30 (0.64)	–0.45 (0.67)
A2	0.1	0.11 (0.04)	0.12 (0.03)	0.01	0.006 (0.006)	0.006 (0.006)	0.2	0.18 (0.104)	0.18 (0.089)	0.5	0.43 (0.208)	0.45 (0.156)	–0.22	–0.08 (0.53)	–0.12 (0.53)
A2	0.01	0.1 (0.03)	0.11 (0.03)	0.01	0.002 (0.003)	0.004 (0.005)	0.2	0.15 (0.089)	0.15 (0.074)	0.5	0.335 (0.259)	0.35 (0.208)	–0.22	–0.01 (0.5)	–0.03 (0.56)
A3	0.1	0.1 (0.01)	0.1 (0.03)	0.01	0.006 (0.006)	0.006 (0.009)	0.5	0.48 (0.089)	0.5 (0.059)	0.8	0.77 (0.319)	0.78 (0.282)	0.24	0.14 (0.59)	0.22 (0.47)
A3	0.01	0.045 (0.01)	0.05 (0.04)	0.01	0.003 (0.004)	0.005 (0.007)	0.5	0.48 (0.059)	0.49 (0.044)	0.8	0.79 (0.311)	0.745 (0.378)	0.24	0.04 (0.37)	0.1 (0.43)
AFA1	0.1	0.07 (0.09)	0.06 (0.08)	0.03	0.01 (0.012)	0.01 (0.007)	0.13	0.09 (0.074)	0.09 (0.074)	0.23	0.195 (0.259)	0.19 (0.248)	0.0	–0.08 (0.77)	–0.12 (0.82)
AFA1n1000	0.1	0.08 (0.11)	0.06 (0.08)	0.03	0.01 (0.007)	0.01 (0.007)	0.13	0.09 (0.059)	0.1 (0.074)	0.23	0.17 (0.222)	0.17 (0.222)	0.0	–0.11 (0.65)	–0.18 (0.71)
AFA2	0.5	0.1 (0.14)	0.05 (0.07)	0.003	0.045 (0.058)	0.045 (0.052)	0.25	0.25 (0.252)	0.38 (0.237)	0.5	0.49 (0.489)	0.545 (0.489)	<0.01	–0.04 (0.8)	0.0 (0.92)

For more details, see Appendix Tables A1 and A5.

Table A8. Estimation of trait-model parameters of DST and DSQ pedigrees (overdominant models)

Model name	P	Estimated median p (MAD)		F_0	Estimated median f_0 (MAD)		F_1	Estimated median f_1 (MAD)		F_2	Estimated median f_2 (MAD)	
		DST	DSQ		DST	DSQ		DST	DSQ		DST	DSQ
U1	0.1	0.12 (0.09)	0.12 (0.09)	0.01	0.008 (0.007)	0.008 (0.004)	0.2	0.09 (0.074)	0.1 (0.059)	0.01	0.25 (0.319)	0.31 (0.267)
U1	0.01	0.1 (0.03)	0.11 (0.04)	0.01	0.003 (0.004)	0.005 (0.006)	0.2	0.14 (0.089)	0.14 (0.074)	0.01	0.31 (0.267)	0.34 (0.208)
U2	0.1	0.11 (0.03)	0.11 (0.03)	0.01	0.005 (0.007)	0.005 (0.007)	0.5	0.5 (0.089)	0.51 (0.044)	0.01	0.07 (0.104)	0.03 (0.044)
U2	0.01	0.045 (0.01)	0.05 (0.03)	0.01	0.002 (0.003)	0.005 (0.007)	0.5	0.49 (0.059)	0.49 (0.044)	0.01	0.63 (0.549)	0.58 (0.563)
U3	0.1	0.11 (0.03)	0.11 (0.03)	0.01	0.006 (0.009)	0.006 (0.009)	0.8	0.81 (0.03)	0.8 (0.03)	0.01	0.025 (0.037)	0.01 (0.015)
U3	0.01	0.02 (0.02)	0.03 (0.03)	0.01	0.008 (0.006)	0.008 (0.006)	0.8	0.8 (0.03)	0.8 (0.03)	0.01	0.89 (0.163)	0.73 (0.4)
U3n1000	0.01	0.02 (0.01)	0.025 (0.01)	0.01	0.008 (0.003)	0.01 (0.003)	0.8	0.8 (0.03)	0.8 (0.015)	0.01	0.86 (0.208)	0.64 (0.43)
U4	0.35	0.36 (0.07)	0.36 (0.09)	0.01	0.003 (0.004)	0.01 (0.015)	0.9	0.9 (0.03)	0.9 (0.015)	0.01	0.001 (0.001)	0.0 (0.0)

For more details, see Appendix Tables A1 and A5.

Table A9. Estimation of trait-model parameters of three-generation (3-G) pedigrees (recessive models)

Model name	P	Estimated median p (MAD)	F_0	Estimated median f_0 (MAD)	F_1	Estimated median f_1 (MAD)	F_2	Estimated median f_2 (MAD)	D	Estimated median d (MAD)
R1	0.1	0.01 (0.01)	0.01	0.001 (0.0003)	0.01	0.001 (0.0007)	0.2	0.11 (0.015)	−1.0	−1.0 (0.01)
R1	0.01	0.08 (0.11)	0.01	0.235 (0.348)	0.01	0.11 (0.163)	0.2	0.32 (0.474)	−1.0	−0.39 (1.14)
R1n1000	0.01	0.075 (0.1)	0.01	0.19 (0.281)	0.01	0.11 (0.163)	0.2	0.28 (0.415)	−1.0	−0.24 (1.21)
R2	0.1	0.09 (0.03)	0.01	0.008 (0.01)	0.01	0.008 (0.003)	0.5	0.51 (0.052)	−1.0	−0.99 (0.04)
R2	0.01	0.06 (0.08)	0.01	0.015 (0.022)	0.01	0.02 (0.03)	0.5	0.41 (0.46)	−1.0	−0.98 (0.27)
R3	0.1	0.09 (0.03)	0.01	0.008 (0.01)	0.01	0.01 (0.003)	0.8	0.8 (0.03)	−1.0	−1.0 (0.03)
R3	0.01	0.05 (0.05)	0.01	0.035 (0.037)	0.01	0.043 (0.033)	0.8	0.76 (0.267)	−1.0	−0.99 (0.06)
AFR1	0.2	0.11 (0.15)	0.04	0.01 (0.015)	0.04	0.015 (0.021)	0.2	0.16 (0.185)	−1.0	−0.94 (0.27)
AFR1n1000	0.2	0.1 (0.13)	0.04	0.01 (0.015)	0.04	0.01 (0.013)	0.2	0.165 (0.185)	−1.0	−0.98 (0.19)
AFR2	0.25	0.23 (0.06)	0.003	0.001 (0.001)	0.05	0.05 (0.015)	0.5	0.51 (0.059)	−0.81	−0.83 (0.07)

For more details, see Appendix Table A1.

Table A10. Estimation of trait-model parameters of three-generation (3-G) pedigrees (dominant models)

Model name	P	Estimated median p (MAD)	F_0	Estimated median f_0 (MAD)	F_1	Estimated Median f_1 (MAD)	F_2	Estimated median f_2 (MAD)	D	Estimated median d (MAD)
D1	0.1	0.08 (0.07)	0.01	0.01 (0.003)	0.2	0.14 (0.074)	0.2	0.19 (0.163)	1.0	0.68 (1.32)
D1	0.01	0.035 (0.01)	0.01	0.004 (0.005)	0.2	0.19 (0.059)	0.2	0.27 (0.163)	1.0	0.33 (0.76)
D1n1000	0.01	0.035 (0.01)	0.01	0.004 (0.005)	0.2	0.19 (0.044)	0.2	0.27 (0.119)	1.0	0.39 (0.62)
D2	0.1	0.09 (0.03)	0.01	0.008 (0.012)	0.5	0.5 (0.044)	0.5	0.48 (0.193)	1.0	0.96 (0.77)
D2	0.01	0.015 (0.01)	0.01	0.01 (0.003)	0.5	0.5 (0.03)	0.5	0.54 (0.563)	1.0	0.42 (0.82)
D3	0.1	0.1 (0.03)	0.01	0.01 (0.015)	0.8	0.8 (0.03)	0.8	0.8 (0.148)	1.0	0.98 (0.39)
D3	0.01	0.015 (0.01)	0.01	0.008 (0.003)	0.8	0.8 (0.015)	0.8	0.84 (0.237)	1.0	0.68 (0.33)
AFD1	0.05	0.11 (0.1)	0.04	0.015 (0.015)	0.2	0.22 (0.104)	0.2	0.22 (0.178)	1.0	0.26 (0.91)
AFD1n1000	0.05	0.11 (0.07)	0.04	0.02 (0.015)	0.2	0.21 (0.096)	0.2	0.21 (0.148)	1.0	0.45 (0.79)
AFD2	0.25	0.18 (0.12)	0.003	0.04 (0.052)	0.5	0.43 (0.074)	0.5	0.43 (0.222)	1.0	1.08 (0.99)

For more details, see Appendix Table A1.

Hum Hered 2016;82:103–139
DOI: 10.1159/000479738

Table A11. Estimation of trait-model parameters of three-generation (3-G) pedigrees (additive models)

Model name	P	Estimated median p (MAD)	F_0	Estimated median f_0 (MAD)	F_1	Estimated median f_1 (MAD)	F_2	Estimated median f_2 (MAD)	D	Estimated median d (MAD)
A1	0.1	0.09 (0.06)	0.01	0.008 (0.01)	0.1	0.1 (0.074)	0.2	0.21 (0.148)	–0.05	–0.09 (0.44)
A1	0.01	0.045 (0.05)	0.01	0.01 (0.0)	0.1	0.1 (0.03)	0.2	0.12 (0.119)	–0.05	0.14 (0.89)
A1n1000	0.01	0.02 (0.02)	0.01	0.01 (0.0)	0.1	0.1 (0.015)	0.2	0.14 (0.133)	–0.05	0.19 (0.9)
A2	0.1	0.08 (0.05)	0.01	0.01 (0.009)	0.2	0.18 (0.089)	0.5	0.52 (0.163)	–0.22	–0.31 (0.25)
A2	0.01	0.035 (0.01)	0.01	0.004 (0.005)	0.2	0.19 (0.074)	0.5	0.34 (0.193)	–0.22	0.04 (0.5)
A3	0.1	0.09 (0.03)	0.01	0.008 (0.012)	0.5	0.5 (0.059)	0.8	0.82 (0.148)	0.24	0.19 (0.27)
A3	0.01	0.015 (0.01)	0.01	0.01 (0.003)	0.5	0.5 (0.044)	0.8	0.765 (0.348)	0.24	0.16 (0.4)
AFA1	0.1	0.09 (0.1)	0.03	0.015 (0.022)	0.13	0.11 (0.089)	0.23	0.25 (0.237)	0.0	–0.13 (0.58)
AFA1n1000	0.1	0.09 (0.07)	0.03	0.02 (0.022)	0.13	0.13 (0.089)	0.23	0.24 (0.193)	0.0	–0.12 (0.44)
AFA2	0.5	0.38 (0.25)	0.003	0.05 (0.074)	0.25	0.23 (0.178)	0.5	0.5 (0.267)	<0.01	–0.09 (0.7)

For more details, see Appendix Table A1.

Table A12. Estimation of trait-model parameters of three-generation (3-G) pedigrees (overdominant models)

Model name	P	Estimated median p (MAD)	F_0	Estimated median f_0 (MAD)	F_1	Estimated median f_1 (MAD)	F_2	Estimated median f_2 (MAD)
U1	0.1	0.12 (0.03)	0.01	0.005 (0.007)	0.2	0.2 (0.074)	0.01	0.04 (0.044)
U1	0.01	0.035 (0.02)	0.01	0.003 (0.004)	0.2	0.19 (0.074)	0.01	0.22 (0.148)
U2	0.1	0.1 (0.01)	0.01	0.008 (0.003)	0.5	0.5 (0.044)	0.01	0.025 (0.037)
U2	0.01	0.015 (0.01)	0.01	0.008 (0.003)	0.5	0.5 (0.044)	0.01	0.13 (0.193)
U3	0.1	0.1 (0.01)	0.01	0.008 (0.003)	0.8	0.8 (0.03)	0.01	0.01 (0.015)
U3	0.01	0.015 (0.01)	0.01	0.008 (0.003)	0.8	0.8 (0.015)	0.01	0.33 (0.489)
U3n1000	0.01	0.015 (0.01)	0.01	0.008 (0.003)	0.8	0.8 (0.015)	0.01	0.28 (0.415)
U4	0.35	0.38 (0.1)	0.01	0.01 (0.007)	0.9	0.9 (0.015)	0.01	0.006 (0.008)

For more details, see Appendix Table A1.

References

1 Ginsburg E, Malkin I, Elston RC: Theoretical Aspects of Pedigree Analysis. Tel Aviv, Ramot Publishing House, 2006.
2 Hasstedt SJ: Pedigree Analysis Software, version 7.1. Salt Lake City, Department of Human Genetics, University of Utah, 2009.
3 S.A.G.E. 6.4. Statistical Analysis for Genetic Epidemiology. 2016. http://darwin.cwru.edu/sage (accessed: October 5, 2017).
4 Thompson EA: Monte Carlo in Genetic Analysis. Technical Report No. 294. Seattle, Department of Statistics, University of Washington, 1995.
5 Thompson EA: Statistical Inferences from Genetic Data on Pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics, vol 6. Beachwood, Institute of Mathematical Statistics, 2000.
6 Kriszt A, Losonczy G, Berta A, Vereb G, Takács L: Segregation analysis suggests that Keratoconus is a complex non-Mendelian disease. Acta Ophthalmol 2014;92:e562–e568.
7 Rao DC: CAT scans, PET scans, and genomic scans. Genet Epidemiol 1998;15:1–18.
8 Elston RC: Methods of linkage analysis – and the assumptions underlying them. Am J Hum Genet 1998;63:931–934.
9 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;44:44–51.
10 Strauch K: MOD-score analysis with simple pedigrees: an overview of likelihood-based linkage methods. Hum Hered 2007;64:192–202.
11 Elston RC: Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. Am J Med Genet 1989;34:487–488.
12 Risch N: Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 1984;36:363–386.
13 Greenberg DA: Linkage analysis assuming a single-locus mode of inheritance for traits determined by two loci: inferring mode of inheritance and estimating penetrance. Genet Epidemiol 1990;7:467–479.
14 Elston RC, Sobel E: Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 1979;31:62–69.
15 Sawyer S: Maximum likelihood estimators for incorrect models, with an application to ascertainment bias for continuous characters. Theor Popul Biol 1990;38:351–366.
16 Ginsburg E, Malkin I, Elston RC: Sampling correction in linkage analysis. Genet Epidemiol 2004;27:87–96.
17 Ewens WJ, Shute NC: A resolution of the ascertainment sampling problem. I. Theory. Theor Popul Biol 1986;30:388–412.
18 Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J: Effects of misspecifying genetic parameters in lod score analysis. Biometrics 1986;42:393–399.
19 Malkin I, Elston RC: Response to letter by Veronica J. Vieland and Susan E. Hodge. Genet Epidemiol 2005;28:286–287.
20 Hodge SE, Vieland VJ: The essence of single ascertainment. Genetics 1996;144:1215–1223.
21 Vieland VJ, Hodge SE: Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. Am J Hum Genet 1995;56:33–43.
22 Vieland VJ, Hodge SE: Ascertainment bias in linkage analysis: comments on Ginsburg et al. Genet Epidemiol 2005;28:283–285.
23 Slager SL, Vieland VJ: Investigating the numerical effects of ascertainment bias in linkage analysis: development of methods and preliminary results. Genet Epidemiol 1997;14:1119–1124.
24 Liang KY, Rathouz PJ, Beaty TH: Determining linkage and mode of inheritance: mode scores and other methods. Genet Epidemiol 1996;13:575–593.
25 Brugger M, Strauch K: Fast linkage analysis with MOD scores using algebraic calculation. Hum Hered 2014;78:179–194.
26 Suarez BK, Rice J, Reich T: The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 1978;42:87–94.
27 Holmans P: Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 1993;52:362–374.
28 Knapp M: A note on linkage analysis with affected sib triplets. Hum Hered 2005;59:21–25.
29 Falls JG, Pulford DJ, Wylie AA, Jirtle RL: Genomic imprinting: implications for human disease. Am J Pathol 1999;154:635–647.
30 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;32:73–83.
31 Dietter J, Mattheisen M, Fürst R, Rüschendorf F, Wienker TF, Strauch K: Linkage analysis using sex-specific recombination fractions with GENEHUNTER-MODSCORE. Bioinformatics 2007;23:64–70.
32 Strauch K. Parametric linkage analysis with automatic optimization of the disease model parameters. Am J Hum Genet 2003;73(suppl 1):A2624.
33 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;66:1945–1957.
34 Knapp M, Strauch K: Affected-sib-pair test for linkage based on constraints for identical-by-descent distributions corresponding to disease models with imprinting. Genet Epidemiol 2004;26:273–285.
35 Strauch K: Gene mapping, imprinting and epigenetics; in Jorde LB, Little PFR, Dunn MJ, Subramaniam S (eds): Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Hoboken, John Wiley & Sons, 2005.
36 Haghighi F, Hodge SE: Likelihood formulation of parent-of-origin effects on segregation analysis, including ascertainment. Am J Hum Genet 2002;70:142–156.
37 Ott J: Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 1989;86:4175–4178.
38 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;71:126–134.
39 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;7:237–243.
40 Xing C, Elston RC: Distribution and magnitude of type I error of model-based multipoint lod scores: implications for multipoint mod scores. Genet Epidemiol 2006;30:447–458.
41 Flaquer A, Strauch K: A comparison of different linkage statistics in small to moderate sized pedigrees with complex diseases. BMC Res Notes 2012;5:411.
42 Shete S, Zhou X: Parametric approach to genomic imprinting analysis with applications to Angelman's syndrome. Hum Hered 2005;59:26–33.
43 Greenberg DA: Inferring mode of inheritance by comparison of lod scores. Am J Med Genet 1989;34:480–486.

Anhang B: Paper IV

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):20180025.

DE GRUYTER Statistical Applications in Genetics and Molecular Biology. 2019; 20180025

Markus Brugger[1,2] / Michael Knapp[3] / Konstantin Strauch[1,2]

Properties and Evaluation of the MOBIT – a novel Linkage-based Test Statistic and Quantification Method for Imprinting

[1] Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Munich, Germany, E-mail: markus.brugger@helmholtz-muenchen.de
[2] Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Ingolstädter Landstraße 1, DE-85764 Neuherberg, Germany, E-mail: markus.brugger@helmholtz-muenchen.de
[3] Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Venusberg-Campus 1, DE-53127 Bonn, Germany

Abstract:
Genomic imprinting is a parent-of-origin effect apparent in an appreciable number of human diseases. We have proposed the new imprinting test statistic MOBIT, which is based on MOD score analysis. We were interested in the properties of the MOBIT concerning its distribution under three hypotheses: (1) $H_{0,\ a}$: no linkage, no imprinting; (2) $H_{0,\ b}$: linkage, no imprinting; (3) H_1: linkage and imprinting. More specifically, we assessed the confounding between imprinting and sex-specific recombination frequencies, which presents a major difficulty in linkage-based testing for imprinting, and evaluated the power of the test. To this end, we have performed a linkage simulation study of affected sib-pairs and a three-generation pedigree with two trait models, many two- and multipoint marker scenarios, three genetic map ratios, two sample sizes, and five imprinting degrees. We also investigated the ability of the MOBIT to quantify the degree of imprinting and applied the MOBIT using a real data example on house dust mite allergy. We further proposed and evaluated two approaches to obtain empiric p values for the MOBIT. Our results showed that twopoint analyses assuming a sex-averaged marker map led to an inflated type I error due to confounding, especially for a larger marker-trait locus distance. When the correct sex-specific marker map was assumed, twopoint analyses have a reduced power to detect imprinting, compared to sex-averaged analyses with an appropriate correction for the inflation of the test statistic. However, confounding was not an issue in multipoint analysis unless the map ratio was extreme and marker spacing was sparse. With multipoint analysis, power as well as the ability to quantify the imprinting degree were almost equally high when a sex-averaged or the correct sex-specific map was used in the analysis. We recommend to obtain empiric p values for the MOBIT using genotype simulations based on the best-fitting nonimprinting model of the real dataset analysis. In addition, an implementation of a method based on the permutation of parental sexes is also available. In summary, we propose to perform multipoint analyses using densely spaced markers to efficiently discover new imprinted loci and to reliably quantify the degree of imprinting.

Keywords: confounding, genomic imprinting, linkage analysis, MOD scores, sex-specific recombination frequencies

DOI: 10.1515/sagmb-2018-0025

1 Introduction

The human genetic map length differs between males and females. This is possibly due to mechanisms closely related to those of genomic imprinting, as regions with sex-specific recombination frequencies often coincide with imprinted ones (Pàldi, Gyapay & Jami, 1995). Genomic imprinting is an epigenetic phenomenon which is present in all viviparous mammals, some plants and, in a wider sense, some insects like the scale insect *Sciara coprophila*, in which it was first described in 1938 (Metz, 1938). As genomic imprinting means the dependence of an individual's liability to develop a disease according to the parental origin of the mutated allele(s), it leads to a deviation from the classic Mendelian assumption of equal contribution of parental genomes to the progeny and is therefore called a parent-of-origin effect (Falls et al., 1999). The degree of genomic imprinting can range from complete inactivation to reduced expression of the respective gene and is established in a time- and tissue-specific manner. Genomic imprinting is caused by DNA and histone modifications without

Markus Brugger is the corresponding author.

changing the nucleotide sequence, i.e. it is epigenetic, and is controlled by imprinting centers (ICs) containing differentially-methylated regions (DMRs) (Lewis & Reik, 2006; Spencer, 2009). Apart from being involved in embryonic development, such that parthenogenetic embryos which arise from genomes of the same parent are hindered to survive to birth, more and more evidence emerges hinting at imprinting being involved in many diseases in the adult. Amongst these are Angelman, Beckwith-Wiedemann, Prader-Willi (Walter & Paulsen, 2003) and Silver-Russell syndrome (Solter, 2006) as well as complex traits like type I diabetes (Bain et al., 1994), atopy (Moffatt & Cookson, 1998), epilepsy (Greenberg et al., 2000), bipolar disorder (Stine et al., 1995) and Alzheimer's disease (Davies, Isles & Wilkinson, 2005). Moreover, the rising incidence of imprinting-associated diseases in children resulting from assisted reproductive technologies (ART) is currently fervently disputed (Wilkins-Haug, 2009). In addition, it has been suggested that imprinting may also play a role in anthropometric traits such as the body mass index (BMI) (Hoggart et al., 2014). The number of imprinted genes seems to be underestimated (Maeda & Hayashizaki, 2006), which is possibly due to the fact that the applied statistical methods deliver inconsistent results and lack power due to variable factors like heterogeneity, penetrance, family and dataset size, and imprinting-mimicking confounders such as sex-specific recombination fractions (Mukhopadhyay & Weeks, 2003; Greenberg et al., 2010).

Imprinting can be tested by linkage analysis methods. Linkage analysis evaluates the co-segregation of genetic marker alleles together with a trait in families. Methods of linkage analysis are commonly distinguished as either being parametric or nonparametric. 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. In the simplest case of a diallelic autosomal trait locus, which is assumed throughout this paper, these parameters are the disease-allele frequency p and 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 is modeled, or the position x of the trait locus in the case of a multipoint analysis. The trait-model parameters can either be pre-specified according to results from previous segregation analyses or maximized along with the recombination fraction in a joint segregation and linkage analysis. For example, a MOD score analysis allows researchers to jointly investigate segregation and linkage (Clerget-Darpoux, Bonaïti-Pellié & Hochez, 1986; Risch, 1984). 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. Our group has implemented the MOD score approach, including a routine to perform simulations under the null hypothesis of no linkage, in the GENEHUNTER-MODSCORE (GHM) software (Brugger & Strauch, 2014; Dietter et al., 2007; Mattheisen et al., 2008; Strauch, 2003), which is based on the GENEHUNTER program (Kruglyak et al., 1996). Its application has led to the identification of a variety of genetic disease loci responsible for congenital heart defects (Flaquer et al., 2013), allergic rhinitis (Kruse et al., 2012), atopic dermatitis (Christensen et al., 2009), bipolar affective disorder (Schumacher et al., 2005), and house dust mite allergy (Kurz et al., 2005). Nonparametric linkage methods have been proposed as an alternative to parametric analysis. These methods promise to avoid trait-model misspecification that may occur when using simple LOD score analyses to map genes responsible for complex traits, for which the mode of inheritance, i.e. trait-model parameters in this case, is usually unknown. Nonparametric methods test if affected pedigree members have more alleles in common than would be expected by chance under the null hypothesis of no linkage. They are often considered to be 'model-free' because they do not rely on explicit assumptions as to the trait-model parameters. However, Knapp, Seuchter, and Baur (1994) have shown that, for samples of affected sib-pairs (ASPs) with the parents' phenotypes unknown or set to unknown, the nonparametric mean test is equivalent to a LOD score analysis under a recessive mode of inheritance, and the possible triangle test proposed by Holmans (1993) is equivalent to a MOD score analysis. In the possible triangle test, the genetic likelihood is expressed in terms of the probabilities z_0, z_1, z_2 that an ASP shares zero, one, or two allele(s) identical-by-descent (IBD) with restrictions to genetically possible models (Holmans, 1993). These allele-sharing probabilities can be expressed as functions of the trait-model parameters f_0, f_1, f_2, p, and θ (Suarez, Rice & Reich, 1978), and hence, the parametric and nonparametric likelihood are identical. Furthermore, Strauch (2007) has shown that the identity of the nonparametric and parametric likelihood holds for any type of pedigree.

In linkage analysis, it is common practice to use sex-averaged genetic maps even if sex-specific differences exist. When using the Kong-and-Cox LOD score (Kong & Cox, 1997) or the MOD score (Risch, 1984; Clerget-Darpoux, Bonaïti-Pellié & Hochez, 1986), this does not change the type I error rate and power of the linkage test, as shown by Fingerlin, Abecasis, and Boehnke (2006) and Dietter et al. (2007), respectively, but only if the ratio of available paternal and maternal genotypes equals 1. In contrast, inflated type I error rate and reduced power is generally observed in the case of a simple parametric multipoint LOD score analysis (Daw, Thompson & Wijsman, 2000). In summary, the direction of the deviation in the type I error rate depends on the sex-specific availability ratio of genotypes, the actual underlying sex-specific map ratio, marker distances,

number of markers, marker information, and sample size (Sieberts & Gudbjartsson, 2005). Therefore it is advisable to adequately model sex-specific recombination frequencies in linkage analysis. The analysis option to include sex-specific recombination frequencies is available in many linkage analysis software packages like Allegro (Gudbjartsson et al. 2000; 2005), Merlin (Abecasis et al., 2002), Superlink (Fishelson & Geiger, 2002), and GHM (Dietter et al., 2007). In contrast to simple LOD score calculations, the power of parametric MOD score analyses is generally less affected by map misspecifications due to the maximization over all model parameters, which effectively serve as nuisance parameters in this case. This holds true for both two- and multipoint analyses. In addition, it has been shown that, especially when analyzing a mixture of different types of pedigrees, the MOD score approach outperforms other linkage methods in terms of power to identify genes with modest effect (Flaquer & Strauch, 2012).

Genomic imprinting introduces an asymmetry between paternal and maternal marker transmission patterns and thus can lead to a substantial loss of power when not taken into account. This directly follows from the statement for standard LOD score analysis that the power to detect linkage is maximal if the analysis model corresponds to the true disease model (Clerget-Darpoux, Bonaïti-Pellié & Hochez, 1986). Relating to linkage tests allowing for imprinting, sex differences in genetic maps and imprinting are confounded in both parametric and nonparametric linkage analyses. It is possible to model imprinting by separately maximizing parametric LOD scores over male and female recombination fractions (θ_{male} and θ_{female}, respectively) (Smalley, 1993), where nonpenetrant cases are explained by fictitious recombinations in the imprinted sex. This results in a correct (uninflated) estimate of the recombination frequency in the nonimprinted sex and an increased LOD score, compared to an analysis with a single recombination fraction for both sexes. A more straightforward approach to model imprinting is to use a four-penetrance formulation distinguishing the heterozygotes according to the parental origin of the disease allele $\mathbf{f} = \{f_0, f_{1\,pat}, f_{1\,mat}, f_2\}$ implemented in the programs GHM and GENEHUNTER-IMPRINTING (Strauch et al., 2000a). Conversely, it is possible to 'model' sex-specific recombination frequencies when the above-mentioned four-penetrance formulation is used in the analysis. Importantly, confounding affects both parametric and nonparametric methods through their relation as outlined by Strauch (2007). In twopoint analyses, when modeling imprinting but not accounting for sex-specific recombination frequencies, imprinting test results can be confounded, leading to an increased type I error rate of the applied test statistic. A true parent-of-origin effect in regions with sex-specific recombination frequencies can then only be declared if the nonimprinted sex has the longer genetic map and shows excess allele-sharing in the analysis (Paterson, Naimark & Petronis, 1999). Generally, the power to detect imprinting is bounded from above by the power to detect linkage (Lemire, 2005). Using MOD scores, a parent-of-origin effect seems likely if the MOD score with four penetrances accounting for imprinting ('IMOD score') is remarkably higher than the MOD score with only three penetrances not accounting for imprinting. Our newly proposed test statistic MOBIT (see Methods section) corresponds to the difference between the two aforementioned scores: $MOBIT = (IMOD\ score) - (MOD\ score)$. When applying sex-averaged recombination frequencies when in fact no imprinting but sex-specific recombination fractions are present, the nonimprinting MOD score is reduced due to an increased number of observed recombinations in the sex with the longer genetic map. In contrast, in an IMOD score analysis, an additional recombination in the sex with the longer genetic map is not modeled as such but the offspring is instead interpreted as a nonpenetrant carrier. This leads to an increased difference between the IMOD and the MOD score, i.e. to confounding. However, when performing multipoint analyses, the disease locus is confined between flanking markers. Hence, the possibility of double recombinations between two adjacent markers is quite low as long as the marker framework is adequately dense (Strauch et al., 2000b). In this case, the linkage analysis no longer loses information due to an increased number of recombinations in the sex having the longer genetic map since linkage to at least one side of the marker framework is preserved, and so it can be expected that the confounding vanishes. Still, in the case of a sparse marker framework and/or large sex-specific map ratios, the probability of double recombinations is non-negligible and confounding might reappear (Strauch et al., 2000b). It has been proposed that confounding is not an issue if map ratios are <5:1 for LOD score analyses allowing for imprinting (Mukhopadhyay & Weeks, 2003) or <10:1 for quantitative trait loci (QTL) LOD score analyses accommodating imprinting (Hanson et al., 2001). Furthermore, marker spacings of <5 cM (Vincent et al., 2006) or <1 cM (Wu, Shete & Amos, 2005) have been proposed to be sufficiently dense to avoid confounding. However, there is no comprehensive and consensual answer to the question to what extent marker spacing, sex-specific map ratios, sample size, and pedigree structure influence confounding in two- and multipoint analyses. This issue is addressed in our extensive simulation study using nuclear families with two affected siblings and extended pedigrees. Three hypotheses (no linkage and no imprinting; linkage and no imprinting; linkage and imprinting) were considered in order to thoroughly investigate the performance of the likelihood-based imprinting test statistic MOBIT and the degree to which imprinting is confounded with sex-specific maps under various ratios. Further, the effect of the sample size was also investigated. In addition, power and the ability of the MOBIT to quantify imprinting were assessed. Finally, the MOBIT was also applied in a real data example on house dust mite allergy to demonstrate the applicability of the MOBIT, for

which the statistical significance was assessed using two alternative simulation/permutation procedures for the calculation of empiric p values.

Throughout this paper, a monogenic dichotomous trait is considered. It should be noted that methods also exist to map imprinted quantitative and ordinal trait loci, see e.g. Shete, Zhou, and Amos (2003) and Feng and Zhang (2008), respectively. Likewise, it is of note that a variety of nonparametric linkage-based imprinting tests have been proposed (Lemire, 2005; Liu et al., 2005; Wu, Shete & Amos, 2005; Vincent et al., 2006). However, these methods either assume independence of parental meioses (Lemire, 2005; Liu et al., 2005; Vincent et al., 2006), are restricted to nuclear families (Wu, Shete & Amos, 2005; Vincent et al., 2006), do not allow for sex-specific recombination fractions (Vincent et al., 2006), do not offer quantification of imprinting (Lemire, 2005; Liu et al., 2005; Wu, Shete & Amos, 2005), or are not explicitly designed for a maximization over the recombination fraction (Lemire, 2005; Liu et al., 2005; Vincent et al., 2006). Correlated meioses can significantly bias imprinting test results if independent parental meioses are assumed in the analysis (Vincent et al., 2006). It can be shown that parental meioses are independent for a multiplicative or a strictly recessive trait model, i.e. a recessive model without phenocopies. A negative correlation of the parental meioses is obtained for additive and dominant trait models and leads to anticonservative imprinting test results. A positive correlation is induced by recessive trait models with phenocopies as well as under- or overdominant trait models and leads to a conservative test. The already existing nonparametric methods, albeit all of them are acknowledged tests for imprinting, are not further investigated in this work due to their above-mentioned properties that render comparisons with the parametric MOBIT hardly feasible.

The paper is structured as follows. In the Methods section, the general framework of the MOBIT is introduced. Then, MOBIT analyses using either a sex-averaged or a sex-specific map for the analysis along with the ability of the MOBIT to quantify imprinting are outlined. At the end of the Methods section, the simulations and analyses of the present study are explained. This is followed by the Results section presenting the results of the simulation study and the real data example on house dust mite allergy. The paper concludes with the Discussion section and a guideline as to how linkage-based imprinting tests should be performed in practice. In addition, the paper includes an Appendix that contains proofs of the asymptotic distribution of the MOBIT, a proof of the identifiability of the marker-trait locus distance in the case of a sex-specific MOD score analysis, and details of a newly developed MOBIT permutation procedure.

2 Methods

2.1 MOBIT – general framework

Generally, in a nonparametric linkage analysis, imprinting can be assessed by looking at the allele-sharing difference between paternal and maternal meioses (Paterson, Naimark & Petronis, 1999). To investigate parent-of-origin effects in ASPs, a nonparametric test has been proposed by Knapp and Strauch (2004), which does not assume independent parental meioses. In that work, the classical Holmans´ possible triangle test statistic $T_{Holmans}$ for ASPs (Holmans, 1993) is extended to the test statistic T_{ILR} which includes four instead of three IBD allele-sharing probabilities by splitting up the probability z_1 of one allele IBD into two probabilities z_1^{pat} and z_1^{mat} according to the parental origin of the allele, such that $z_1 = z_1^{pat} + z_1^{mat}$. Regarding the parameter space, this leads to an extension of the possible triangle to a tetrahedron which accounts for disease models with $z_1^{pat} \neq z_1^{mat}$ (see Figure 1). When analyzing ASPs with parental phenotypes set to unknown and employing a sex-averaged marker map, the parametric MOD score is equivalent to the nonparametric $T_{Holmans}$ (Knapp, Seuchter & Baur, 1994) and the MOD score with four penetrances accounting for imprinting (IMOD score) is equivalent to the nonparametric T_{ILR} (Knapp & Strauch, 2004). Going beyond a test for linkage adaptive to imprinting, the MOD and the IMOD score can be combined to test the null hypothesis of linkage but no imprinting by evaluation of the difference of the respective maximized log-likelihood ratios with and without distinguishing the heterozygotes, which is given by

$$
\begin{aligned}
MOBIT &= (IMOD\ score) - (MOD\ score) \\
&= \log_{10}\frac{L(\hat{z}_0,\ \hat{z}_1^{pat},\ \hat{z}_1^{mat},\ \hat{z}_2)}{L(\frac{1}{4},\ \frac{1}{4},\ \frac{1}{4},\ \frac{1}{4})} - \log_{10}\frac{L(\hat{z}_0,\ \hat{z}_1,\ \hat{z}_2)}{L(\frac{1}{4},\ \frac{1}{2},\ \frac{1}{4})} \\
&= \log_{10}(L(\hat{z}_0,\ \hat{z}_1^{pat},\ \hat{z}_1^{mat},\ \hat{z}_2)) - \log_{10}\left(L\left(\frac{1}{4},\ \frac{1}{4},\ \frac{1}{4},\ \frac{1}{4}\right)\right) \\
&\quad - \log_{10}(L(\hat{z}_0,\ \hat{z}_1,\ \hat{z}_2)) + \log_{10}\left(L\left(\frac{1}{4},\ \frac{1}{2},\ \frac{1}{4}\right)\right) \\
&= \log_{10}\frac{L(\hat{z}_0,\ \hat{z}_1^{pat},\ \hat{z}_1^{mat},\ \hat{z}_2)}{L(\hat{z}_0,\ \hat{z}_1,\ \hat{z}_2)}
\end{aligned}
\tag{1}
$$

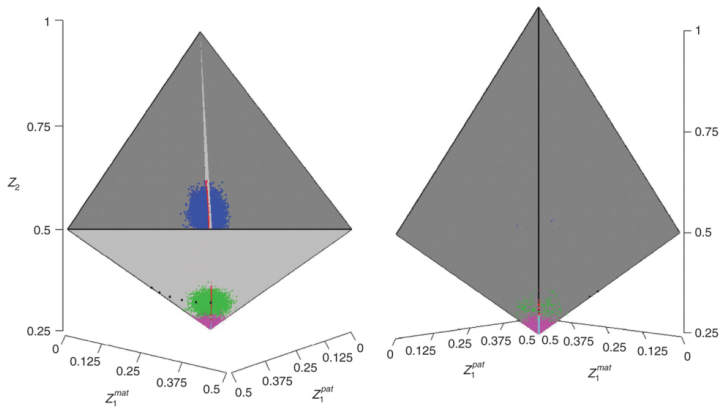


Figure 1: Tetrahedron **T** defining the set of allele-sharing probabilities $\mathbf{z} = \left(z_0,\ z_1^{pat},\ z_1^{mat},\ z_2\right)$ for affected sib-pairs (ASPs) that correspond to meaningful genetic models, as a subset of the larger set of all possible allele-sharing probabilities represented by the outer cube.
Between the left half-tetrahedron $\mathbf{T}_{z_1^{pat}>z_1^{mat}}$ and the right half-tetrahedron $\mathbf{T}_{z_1^{pat}<z_1^{mat}}$ (both in dark-grey) lies the sagittal plane (light-grey), which corresponds to the possible triangle in the nonimprinting parameter space $(z_0,\ z_1,\ z_2)$, dividing **T** into two halves. The left subfigure depicts **T** with the front plane (light-grey) facing towards the beholder, whereas in the right subfigure **T** is turned around, such that the nonimprinting parameter space $(z_0,\ z_1,\ z_2)$can be seen as a black line dividing **T** into two halves. Black bullet points on the front triangle (also in light-grey) in the left subfigure correspond to the points in terms of allele sharing that were used to simulate the additive trait model. Hence, the black bullet in the middle of the frontern triangle corresponds to penetrances $f_0 = 0.03,\ f_1 = 0.13,\ f_2 = 0.23$, and disease allele frequency $p = 0.1$. The black bullets on the left side of the front triangle (left subfigure) correspond to imprinting models, such that the farther left a bullet point lies, the higher is its corresponding imprinting index ($\mathbf{I} = \{0.2;\ 0.4;\ 0.6;\ 0.8;\ 1\}$). Green dots on the light-grey front plane represent maximum likelihood estimates (MLEs) of a MOD score analysis that takes imprinting into account (IMOD score) using 10,000 replicates of 600 ASPs and a fully informative marker with recombination fraction $\theta = 0$ and the additive trait model. The corresponding nonimprinting MOD score MLEs (red dots) lie in the middle of the light-grey front plane, which corresponds to the upper edge of the possible triangle, i.e. the nonimprinting parameter space $(z_0,\ z_1,\ z_2)$. Pink and turquoise dots correspond to IMOD and MOD score MLEs under no linkage, respectively. With regard to the simulated recessive trait model with $f_0 = 0.05,\ f_1 = 0.05,\ f_2 = 0.9$, and disease allele frequency $p = 0.2$, blue and red dots in the upper part of the left subfigure correspond to the IMOD and MOD score MLEs, respectively.

where $\hat{z}_{i=0,\ 1,\ 2}$ denote the maximum likelihood estimators (MLEs) restricted to the possible set of allele-sharing probabilities **z** defined by the possible triangle (Δ) for the $T_{Holmans}$ with $\mathbf{z}_\Delta = \{(z_0,\ z_1,\ z_2) :\ z_1 \le 0.5,\ 2 \cdot z_0 \le z_1,\ z_0 \ge 0,\ z_1 \ge 0,\ z_2 \ge 0,\ z_2 = 1 - z_0 - z_1\}$ and the tetrahedron (**T**) for the T_{ILR} with $\mathbf{z}_\mathbf{T} = \{(z_0,\ z_1^{pat},\ z_1^{mat},\ z_2) :\ z_1^{pat} + z_1^{mat} \le 0.5,\ z_0 \le z_1^{pat} \le z_2,\ z_0 \le z_1^{mat} \le z_2,\ z_2 \ge 0,\ z_1^{pat} \ge 0,\ z_1^{mat} \ge 0,\ z_0 \ge 0,\ z_0 = 1 - z_2 - z_1^{pat} - z_1^{mat}\}$. The equivalence of the likelihoods under the null hypothesis of no linkage $L(\frac{1}{4},\ \frac{1}{4},\ \frac{1}{4},\ \frac{1}{4})$ and $L(\frac{1}{4},\ \frac{1}{2},\ \frac{1}{4})$ is shown in Appendix A.1 of the paper. In their Discussion section, Knapp and Strauch (2004) have proposed a nonparametric imprinting test for ASPs that is (up to a factor of 2 $ln(10)$) identical to the MOBIT in the context of ASPs. Denoting this nonparametric imprinting test by T_{Impr}, we can therefore write

$$MOBIT = \frac{1}{2\ln(10)} \cdot 2\ln\frac{L(\hat{z}_0,\ \hat{z}_1^{pat},\ \hat{z}_1^{mat},\ \hat{z}_2)}{L(\hat{z}_0,\ \hat{z}_1,\ \hat{z}_2)} = \frac{T_{Impr}}{2\ln(10)} \tag{2}$$

The null distribution of the T_{Impr} depends on the true underlying $\mathbf{z} \in H_0 : z_1^{pat} = z_1^{mat}$ which is either an interior point of $\mathbf{T}$ or lying on its boundary. If $\mathbf{z} \in H_0 : z_1^{pat} = z_1^{mat}$ is an interior point of $\mathbf{T}$, standard theory predicts that the asymptotic distribution of the T_{Impr} is χ^2 with 1 df, although the proximity to the boundary of $\mathbf{T}$ may affect the quality of the asymptotic approximation. If $\mathbf{z}$ is a point on the boundary of $\mathbf{T}$, the quantiles of the asymptotic distribution can be smaller than for a χ^2 distribution with 1 df (Knapp & Strauch, 2004). Due to the equality, when analyzing ASPs, the same properties hold true for the MOBIT. For the point of no linkage $\mathbf{z} = \left(\frac{1}{4},\ \frac{1}{4},\ \frac{1}{4},\ \frac{1}{4}\right)$, the MOBIT follows a mixture of distributions that includes non-χ^2 components (see also Self and Liang (1987), case 8, pp. 608–609). It is noteworthy that the equality of a parametric MOD score analysis with a corresponding allele-sharing-based test statistic holds in general for arbitrary pedigree structures (Strauch, 2007). However, the corresponding allele-sharing configurations have only been formulated for unilineal affected relative-pairs (ARPs), ASPs, and affected sib-triplets (ASTs) so far (Knapp, 2005; Strauch, 2007). The allele-sharing parameter spaces for larger pedigrees involve a larger number of dimensions, and the corresponding restrictions for genetically possible models are expected to have a more complicated form (Knapp, 2005; Strauch, 2007). Even for ASTs, the parameter restrictions are unknown so far, which precludes the construction of an allele-sharing-based test (Knapp, 2005). Therefore, the null distribution of the MOBIT cannot be analytically derived from an equivalent nonparametric test for larger pedigrees. However, given the equality of the nonimprinting and imprinting likelihoods under the null hypothesis of no linkage and no imprinting for any type of pedigree (Appendix A.2), the distribution is expected to be χ^2 with 1 df, because standard likelihood ratio technique predicts that the number of degrees of freedom is equal to the difference between the maximized parameters in the numerator and the denominator, which equals $4 - 3 = 1$ in our case, which represents the case of two nested composite hypotheses (Wilks, 1938) (see also Appendix A.3).

2.2 Sex-averaged MOBIT analysis

Assuming sex-averaged recombination fractions, a MOD score analysis using ASPs is equivalent to $T_{Holmans}$, as shown by Knapp, Seuchter, and Baur (1994), and thus explores the same parameter space of $\mathbf{z}_\triangle$ defined by the possible triangle (whole triangle, light-grey sagittal plane in Figure 1). In particular, a given combination of penetrances $f_0,\ f_1,\ f_2$, disease allele frequency p, and recombination frequency θ in a parametric analysis corresponds to a certain point $\mathbf{z}$ within the possible triangle in the nonparametric context. How certain sets of parametric trait-model parameters for ASPs with or without imprinting translate into allele-sharing probabilities can be found in Knapp and Strauch (2004) and Suarez, Rice, and Reich (1978), respectively. Accordingly, the maximization of the parametric likelihood ratio over the four trait-model parameters and θ in the MOD score approach corresponds to maximizing the nonparametric allele-sharing parameters over the two-dimensional plane of the possible triangle. The IMOD score analysis, on the other hand, explores the whole tetrahedron $\mathbf{T}$ including points for which $z_1^{pat} \neq z_1^{mat}$ and is equivalent to the T_{ILR}. In this case, the parametric likelihood ratio is maximized over the disease allele frequency p, the recombination frequency θ, and four penetrance parameters $f_0,\ f_{1\,pat},\ f_{1\,mat},\ f_2$ distinguishing the heterozygotes, where $f_{1\,pat} \neq f_{1\,mat}$ is equivalent to excess paternal or maternal allele-sharing, i.e. $z_1^{pat} \neq z_1^{mat}$. Throughout this paper, without loss of generality, $\theta_{male} \leq \theta_{female}$, is assumed. If no imprinting is present, this corresponds to points in terms of allele-sharing within the left half-tetrahedron ($\mathbf{T}_{\mathbf{z}_1^{\mathrm{pat}} \geq \mathbf{z}_1^{\mathrm{mat}}}$) in Figure 1.

2.3 Sex-specific MOBIT analysis

The parameter space explored by the maximization of the imprinting likelihood remains unchanged when using a sex-specific marker map, i.e. it is still the whole tetrahedron $\mathbf{T}$ for ASPs. However, the maximization of the nonimprinting likelihood is now enabled to explore parameter space beyond the possible triangle, according to the sex with the longer genetic map. For a given set of trait-model parameters $f_0,\ f_1,\ f_2$, and p, the parametric nonimprinting maximization starts with $\theta_{male} = \theta_{female} = 0$ at the corresponding point on the possible triangle. This holds for a sex-specific as well as for a sex-averaged genetic map. Then, in the sex-specific case, the recombination fraction is varied according to the specified genetic map ratio, i.e. $0 < \theta_{male} < \theta_{female} < \frac{1}{2}$ in the analysis. This leads to points $\mathbf{z}$ within the left half-tetrahedron $\mathbf{T}_{\mathbf{z}_1^{\mathrm{pat}} > \mathbf{z}_1^{\mathrm{mat}}}$ outside the sagittal plane $\mathbf{z}_\triangle$, i.e. the

possible triangle. The maximization over the recombination frequency continues until the point of no linkage $\theta_{male} = \theta_{female} = \frac{1}{2}$ is reached, which corresponds to $\mathbf{z} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. Hence, each assumed pair of recombination fractions $\theta_{male} \leq \theta_{female}$ corresponds to a point within $\mathbf{T}_{z_1^{pat} \geq z_1^{mat}}$. These points join to a maximization curve over the recombination fraction for a given set of disease model parameters and a fixed female/male map ratio as illustrated in Figure 2 (black line with grey diamonds). The maximization is then continued over arbitrary combinations of trait-model parameters. The range and extent of the parameter space explored by the nonimprinting maximization within $\mathbf{T}_{z_1^{pat} > z_1^{mat}}$ depend on (1) the assumed genetic map ratio and (2) the inter-marker distance in the case of a multipoint analysis. The more extreme the assumed map ratio, the farther reach such maximization curves into $\mathbf{T}_{z_1^{pat} > z_1^{mat}}$. The same holds for larger inter-marker distances. The point of maximum maternal imprinting $\mathbf{z} = \left(0, \frac{1}{2}, 0, \frac{1}{2}\right)$ is only reached with an infinite female/male map ratio. Interestingly, given a truly underlying nonimprinting disease model with $f_{1\,pat} = f_{1\,mat}$, a genetic map ratio less or larger than 1, and $\theta > 0$ between marker and trait locus, the respective marker-trait locus distance is identifiable in a MOD score analysis using ASPs (see Appendix B for a proof). Using a sex-specific map in the analysis should avoid confounding between genomic imprinting and sex-specific recombination fractions, such that the type I error rate of the imprinting test does not exceed its nominal level. However, the question arises, whether the power of the test can be reduced due to a confounding between the genetic position and the trait-model parameters in the maximization of the likelihoods in equation (1). That is because imprinting can be "modeled" by separately maximizing the likelihood over male and female recombination fractions (θ_{male} and θ_{female}, respectively) (Smalley, 1993), which is already done in the nonimprinting likelihood, just as well as by using the four-penetrance formulation distinguishing the heterozygotes according to the parental origin of the disease allele $\mathbf{f} = \{f_0, f_{1\,pat}, f_{1\,mat}, f_2\}$ (Strauch et al., 2000a) in the imprinting likelihood. However, this effect is expected to become negligible in the case of a densely spaced marker framework.

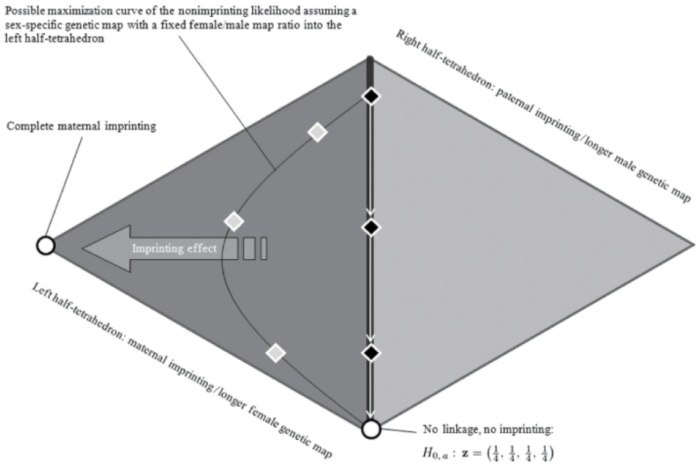


Figure 2: A graphical representation of a transversal cut through the tetrahedron **T** is shown to illustrate the nonimprinting likelihood maximization in terms of the allele-sharing parameters $\mathbf{z}=\left(z_0, z_1^{pat}, z_1^{mat}, z_2\right)$ within **T** in an analysis using affected sib-pairs (ASPs). The effect of a female/male genetic map ratio larger than 1 and maternal imprinting is considered. For a given set of penetrances f_0, f_1, f_2, and the disease allele frequency p, the sex-averaged nonimprinting maximization over the recombination frequency θ starts on the corresponding point **z** on the possible triangle (black middle line) assuming $\theta = 0$ (upper black diamond). The recombination fraction is then gradually increased, leading to a curve within the possible triangle (white arrows), until **z** reaches $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ for $\theta = \frac{1}{2}$, i.e. no linkage. In an analysis assuming a sex-specific map, the maximization over θ starts at the same initial point **z** on the possible triangle as in the sex-averaged analysis, given the same set of disease model parameters and $\theta = 0$ (upper black diamond). However, the recombination fraction is now varied according to the genetic map ratio, i.e. $\frac{\theta_{female}}{\theta_{male}} > 1$, which leads to explored points **z** along a curve within the left half-tetrahedron $\mathbf{T}_{\mathbf{z_1^{pat}}>\mathbf{z_1^{mat}}}$ (black solid curve and grey diamonds). At $\theta = \frac{1}{2}$, **z** again reaches $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. The extent to which a sex-specific nonimprinting maximization explores parameter space within $\mathbf{T}_{\mathbf{z_1^{pat}}>\mathbf{z_1^{mat}}}$, i.e. the outreach into the left half-tetrahedron, is increased by a larger female/male genetic map ratio. The point of complete maternal imprinting $\left(0, \frac{1}{2}, 0, \frac{1}{2}\right)$ is only reached for an infinite female/male map ratio. In addition, maternal imprinting causes **z** to be shifted farther left (large grey arrow) so that it no longer lies within the maximization scope of a sex-specific MOD score analysis not allowing for imprinting. These points are then exclusively reached by an IMOD score analysis. Yet, if **z** already lies very near to the boundary of **T** due to large differences in sex-specific recombination fractions (say, **z** corresponds to the lower grey diamond), it will hardly be shifted farther left by maternal imprinting. Hence, the sex-specific nonimprinting MOD score analysis is almost equivalent to the IMOD score analysis with respect to the explored parameter space, which leads to a lower power of the MOBIT, especially when maternal imprinting is incomplete. All these conclusions equally apply to female/male map ratios smaller than 1 and paternal imprinting.

2.4 Quantifying genomic imprinting

Using the MOBIT, imprinting can be quantified by looking at the imprinting index $\boldsymbol{I}$ (Strauch, 2005) calculated from the estimated penetrances at the assumed disease locus showing the highest evidence for imprinting. The imprinting index $\mathbf{I} = \frac{f_{1\ pat} - f_{1\ mat}}{f_2 - f_0}$ equals the difference between the two heterozygote penetrances, normalized by the difference of the homozygote penetrances in order to properly take the case of a nonzero phenocopy rate or reduced penetrance into account. The question as to what extent trait-model parameters can be estimated in a MOD score analysis cannot be answered comprehensively in this paper. However, trait-model parameters can in principle be estimated by a MOD score analysis (Elston, 1989), which had been outlined in the context of the ascertainment-assumption-free method (Shute & Ewens, 1988). However, the identifiability of the trait-model parameters f_0, $f_{1\,pat}$, $f_{1\,mat}$, f_2, p, and θ depends on the number of allele-sharing classes in the dataset. In the case of ASPs, the allele-sharing classes are z_0, z_1^{pat}, z_1^{mat}, and z_2 when taking imprinting into account. Hence, as there are only $4 - 1 = 3$ free parameters that can be estimated from ASP data, there will be many sets of $(f_0, f_{1\,pat}, f_{1\,mat}, f_2, p, \text{and } \theta\theta)$ that correspond to a particular estimated $\left(z_0, z_1^{pat}, z_1^{mat}, \text{and } z_2\right)$. With larger pedigrees, and hence more allele-sharing classes, the degree to which the trait-model parameters can be correctly determined should be higher. It is of note, however, that for any type of affecteds-only analysis, the absolute values of penetrances cannot be determined, because multiplication of all penetrances by the same factor does not change the result. In order to investigate the ability of our newly proposed MOBIT test to quantify imprinting, we compared the estimated imprinting degrees in terms of the imprinting index $\boldsymbol{I}$ with the simulated values when either using a sex-specific or a sex-averaged map in the analysis.

2.5 Simulation and analysis

The families under study were samples of either affected sib-pairs (ASPs) or extended pedigrees with three generations (3-G) (see Figure 3). A diallelic disease locus causing a dichotomous trait, with parameters that should reflect the characteristics of complex disorders, was chosen for the simulations. In order to ensure that the power to detect imprinting is sufficiently high with replicates generated under the selected parameter set and that the computations are still feasible, power calculations for the linkage test with ASPs were done using the *R* package *powerpkg* (Weeks, 2010) under various parameter sets prior to performing the simulations. The parental trait phenotypes were set to unknown, and a fully informative marker in complete linkage, i.e. $\theta = 0$, with the disease locus was simulated. An additive single-locus disease model with penetrances $\{f_0, f_1, f_2\} = \{0.03,\ 0.13,\ 0.23\}$ and disease allele frequency $p = 0.1$ was chosen, which leads to a power to detect linkage of approximately 80% using a sample size of 600 ASP families (type I error rate: 10^{-4}). The respective sample sizes for the 3-G pedigrees were derived by initial test simulations. In particular, given 3-G pedigrees, sample size, a fully informative marker, and a type I error rate of 10^{-4}, the critical value for the linkage test was determined under the null hypothesis of no linkage, and power was assessed by simulating completely linked replicates.

This led to a sample size of 65 3-G pedigrees for the additive trait model. To evaluate the effect of sample size for confounding, we additionally analyzed samples consisting of 400 ASPs and 45 3-G pedigrees, which corresponds to a power of 50% to detect linkage. It is of note that the additive trait model lies on the boundary of the parameter space of ASPs, which affects the null distribution of the MOBIT. Hence, a recessive trait model with $p = 0.2$ and $\{f_0, f_1, f_2\} = \{0.05,\ 0.05,\ 0.9\}$, which lies in the interior of the parameter space of ASPs, was also considered to numerically verify the χ^2_1 distribution of the MOBIT in the case of no boundary condition. The sample size used for the simulations for the recessive trait model was chosen to be the same as for the additive trait model. This way, the degrees of confounding of the imprinting tests are based on the same number of meioses for both trait models and can thus fairly be compared.

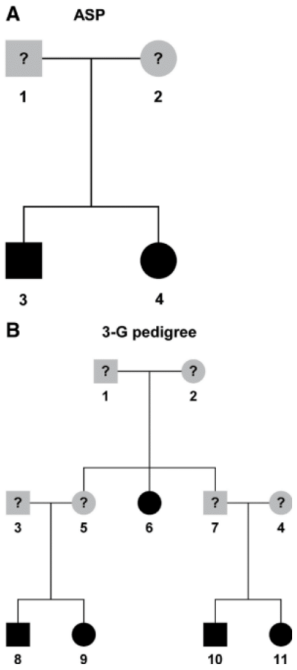


Figure 3: (A) Affected sib-pair (ASP) (B) Three-generation pedigree (3-G).
Pedigrees used for the simulations. ASP: affected sib-pair; 3-G: three-generation pedigree; ?: unknown phenotype; filled symbols: affected; empty symbols: unaffected.

Generation of genotype data with or without imprinting effects and conditional on affected offspring (with parental phenotypes assumed to be unknown) were either carried out by SLINK (Ott, 1989; Schäffer et al., 2011; Weeks et al., 1990) or by its imprinting extension SLINK Imprinting (Shete & Zhou, 2005). The simulation algorithm calculates the probability distribution of genotypes $\mathbf{g} = g_1, g_2, ..., g_n$ conditional on the phenotype values $\mathbf{x} = x_1, x_2, ..., x_n$ 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: $P(\mathbf{g}|\mathbf{x}) = P(g_1|\mathbf{x})P(g_2|g_1, \mathbf{x})P(g_3|g_1, g_2, \mathbf{x})...$ The calculation time of this algorithm increases linearly with additional family members, but exponentially with the number of markers. In order to avoid prohibitive computation times when simulating several markers, a two-step algorithm developed by Lemire (2006) was employed, which exploits the ability of conditional simulations by SLINK or SLINK Imprinting, respectively, and uses a gene dropping algorithm implemented in the SLINK utility program SUP (Lemire, 2006; Schäffer et al., 2011) to quickly generate a large number of markers. This procedure starts with the generation of the disease locus genotypes and trait values by SLINK or SLINK Imprinting, respectively, where a fully informative marker in linkage equilibrium (LE) with the disease locus and $\theta = 0$ is simulated to mark the path of inheritance of the disease; it is therefore called the 'descent marker'. Using SUP, it is then possible to simulate single marker alleles or haplotypes for several markers along the generations allowing for sex-specific recombination frequencies. Four scenarios were simulated: (1) a fully informative microsatellite marker with $\theta = 0$ between the marker and the trait locus; (2) a fully informative microsatellite marker with two nonzero genetic distances to the disease locus; (3) four microsatellites with four equifrequent alleles corresponding to a mean heterozygosity (HET) of 0.75,

with the disease locus placed halfway between marker 2 and 3 at various genetic distances; (4) a typical array of 40 diallelic single-nucleotide polymorphisms (SNPs), each SNP with a minor allele frequency (MAF) of 0.15 corresponding to a mean HET of 0.25, spaced 0.32 cM from each other and the disease locus placed halfway between SNPs 20 and 21. The genetic distances between the marker(s) and the disease locus in the single and four-marker scenarios (2) and (3) were 0.5 and 5 cM, respectively. The former distance implies a marker spacing of 1 cM, which may be used for fine-mapping, and the latter distance corresponds to a sparse marker spacing of 10 cM. Besides sex-equal recombination frequencies, two female/male map ratios were assumed. A map ratio of 7:3 was simulated as a possible value for an imprinted region like the one on chromosome 11p13 around the *WT*1 gene showing a ratio of about 2.1:1 (Pàldi, Gyapay & Jami, 1995), whereas a map ratio of about 9:1 can be found in the pseudoautosomal regions PAR1 and PAR2 on the X and Y chromosome (Flaquer, Fischer & Wienker, 2009; Matise et al., 2007). To evaluate the size and the power of the MOBIT, simulations under three hypotheses were analyzed: $H_{0,\,a}$: no linkage, no imprinting, $H_{0,\,b}$: linkage, no imprinting, and H_1: linkage, imprinting. For the latter, five degrees of maternal imprinting were assumed, corresponding to imprinting indices of $\mathbf{I} = \{0.2;\ 0.4;\ 0.6;\ 0.8;\ 1\}$ as defined by Strauch (2005) and given above. Additionally, paternal imprinting, corresponding to $\mathbf{I} = \{-0.2; -0.4; -0.6; -0.8; -1\}$, was also considered, but only for the scenarios with 5 cM between marker and trait locus. This is because these scenarios correspond to the case in which imprinting occurs in the sex with the shorter genetic map, so that the impact on power due to sex-specific recombination fractions and confounding should be especially relevant in the case of large marker-trait locus distances. In the case of the simulated additive trait model, the penetrances were $f_0 = 0.03$ and $f_2 = 0.23$ and the average of $f_{1\,pat}$ and $f_{1\,mat}$ was 0.13, corresponding to the nonimprinting case. An overview of the simulation scenarios can be found in Table 1. It should be noted that the concepts of dominance, recessivity, or multiplicativity do not make sense under imprinting conditions (Strauch et al., 2000a). More specifically, a recessive trait model like the one we used for the simulations has no corresponding 'recessive' imprinting model within the diamond of inheritance (DOI) (Strauch, 2005). Instead, with an increasing degree of imprinting, the model approaches the horizontal axis of the DOI $\left(\frac{f_{1,\,pat}+f_{1,\,mat}}{2} = \frac{f_0+f_2}{2}\right)$, which is closer to additive than dominant or recessive models. A similar relation holds for the corresponding points in terms of allele-sharing. Therefore, simulations under H_1 were only done for the additive disease model, which lies on the horizontal axis of the DOI, and, for reasons of conciseness, only for the twopoint scenarios (1) and (2) as well as the SNPs scenario (4). The number of replicates was set to 10,000, except for the scenarios under H_1 with 5,000 replicates.

Table 1: Overview of the simulated scenarios to investigate confounding between sex-specific recombination fractions and genomic imprinting. SNP: single nucleotide polymorphism; ASP: affected sib-pair; 3-G pedigree: three-generation pedigree. *Only for the twopoint scenarios with 5 cM marker-trait locus distance.

Simulations under: $H_{0,\,a}$: No linkage, no imprinting; $H_{0,\,b}$: Linkage, no imprinting; H_1: Linkage, imprinting			
	Map ratio (female : male)		
Distance between marker(s) and disease locus (sex-averaged)	1:1	7:3	9:1
0 cM ($\theta = 0$)	1 marker	1 marker	1 marker
0.5 cM ($\theta \approx 0.005$)	1 or 4 marker(s)	1 or 4 marker(s)	1 or 4 marker(s)
5 cM ($\theta \approx 0.048$)	1 or 4 marker(s)	1 or 4 marker(s)	1 or 4 marker(s)
0.16 cM ($\theta \approx 0.0016$)	40 SNPs	40 SNPs	40 SNPs
Recombination fraction θ based on Haldane map function			
4 markers: disease locus halfway between markers 2 and 3 with a marker spacing of 1 and 10 cM, respectively.			
40 SNPs: halfway between markers 20 and 21 with a marker spacing of 0.32 cM.			
Segregation of additive trait simulated with penetrances $\{f_0, f_1, f_2\} = \{0.03,\ 0.13,\ 0.23\}$ and disease allele frequency $p = 0.1$			
Segregation of recessive trait simulated with penetrances $\{f_0, f_1, f_2\} = \{0.05,\ 0.05,\ 0.90\}$ and disease allele frequency $p = 0.2$			
Pedigree type	Sample size 1	Sample size 2	
ASPs	600	400	
3-G pedigrees	65	45	
Maternal imprinting simulations for additive trait model step-wise with $\mathbf{I} = \frac{f_{1\,pat}-f_{1\,mat}}{f_2-f_0} = (0.2;\ 0.4;\ 0.6;\ 0.8;\ 1)$			
Paternal imprinting simulations* for additive trait model with $\mathbf{I} = \frac{f_{1\,pat}-f_{1\,mat}}{f_2-f_0} = (-0.2; -0.4; -0.6; -0.8; -1)$			

Each simulated replicate was subsequently analyzed by the GHM program, in which both the MOD score (option 'imprinting off') and the IMOD score (option 'imprinting on') were calculated in a single program

run. Technically, the calculation of the MOBIT is realized by applying the GHM option 'modcalc global', by which the maximum of the LOD score over all assumed disease locus positions along the marker map is determined for each trait model, and this maximum is then maximized over different trait models. Further, the options 'maximization dense', 'penetrance restriction off', and 'allfreq restriction off' were used in the analysis. In terms of allele-sharing probabilities, the MOBIT compares the MLE of $\mathbf{z_T}$ (imprinting) with the one of $\mathbf{z}_\triangle$ (no imprinting) for ASPs, allowing for different disease locus positions in the numerator and the denominator of the likelihood ratio in equation (1) at which the MLE is calculated. Regarding the twopoint scenarios, the MOBIT was evaluated at the marker locus and at 100 equally spaced genetic positions lying up to 100 cM away from the marker locus. In the multipoint scenarios, the MOBIT was calculated with the putative trait locus positioned directly at the markers and halfway between them. The maximization was done over the aforementioned set of genetic positions x of the putative trait locus or θ in the case of a twopoint scenario as well as over all trait-model parameters, i.e. the disease allele frequency p and the penetrances $f_0,\ f_1,\ f_2$ without imprinting or $f_0,\ f_{1\,pat},\ f_{1\,mat},\ f_2$ with imprinting taken into account. All replicates were analyzed using a sex-averaged map and a sex-specific map with the same genetic map ratio as used for the simulation. The imprinting test statistic *MOBIT* = (*IMOD score*) − (*MOD score*) was calculated and its empiric distributions under all three hypotheses were determined assuming either a sex-averaged or a sex-specific map for the analysis. Finally, the 95% and 99% quantiles of the empiric MOBIT distributions under the $H_{0,\,a}$ and $H_{0,\,b}$ hypotheses were obtained. Empiric quantiles were calculated according to the formula $\mathrm{p} = \frac{k}{n}$, n being the total number of replicates and k the number of ordered replicates. The test statistic of the k-th ordered replicate, which corresponds to a p value with $1 - \mathrm{p}$ equal to 95% or 99% , defines the respective empiric quantile. Type I error rates using a nominal 1% and 5% significance level were calculated using the assumed asymptotic distribution of χ^2 with 1 df. Power was measured as the proportion of replicates simulated under H_1: linkage and imprinting that exceeded the respective empiric 95% quantile determined under $H_{0,\,b}$, thus ruling out inflated type I error rates due to confounding of imprinting with sex-specific recombination frequencies. In addition, we assessed the performance of the MOBIT with respect to its ability to correctly estimate the imprinting index in the scenarios used for the power calculations. We also looked for differences in estimation accuracy between sex-averaged and sex-specific analyses, maternal and paternal imprinting, and the two pedigree types.

2.6 Real data example on house dust mite allergy

Due to the dependence of the asymptotic properties of the MOBIT on the truly underlying point in terms of allele-sharing within the corresponding parameter space, p values obtained from a χ^2_1 distribution might lead to false positive or false negative test results in practice. Therefore, one might simulate the MOBIT distribution under the null hypothesis of linkage, but no imprinting, such that the **b**est-**f**itting **n**onimprinting **m**odel, including the recombination fraction or the genetic position of the disease, is used for the simulation of genotype data (method *bfnm*), which can be done *ab initio* using the software packages SLINK (Ott, 1989; Schäffer et al., 2011; Weeks et al., 1990) and SUP (Lemire, 2006; Schäffer et al., 2011). Method *bfnm* is similar to the simulation approach of the above-described main simulation study using a truly underlying hypothesis, which, however, is unknown in the general case. An alternative approach, similar to the methods proposed by Dong et al. (2005) and Whittaker et al. (2003), might be to obtain a p value by the use of a permutation procedure based on the randomization of the origin of parental alleles in offspring of every nuclear family within a given pedigree (method *perm*). Hence, we have developed a permutation procedure for the MOBIT to obtain empiric p values and implemented it in the GHM software package. Such a procedure effectively isolates the imprinting effect from overall evidence of linkage (Dong et al., 2005). In contrast to method *bfnm*, the null hypothesis of the newly implemented MOBIT-based *perm* procedure corresponds to an imprinting effect with expectation value 0, conditional on the linkage information of the real dataset. As a consequence, application of method *perm* can lead to a completely different sample space compared to method *bfnm*, and can hence lead to different quantiles and p values (see also Appendix C). Importantly, the power of a permutation test is restricted by the sample size, which needs to be sufficiently large to obtain a fine-grained permutation distribution. More details as to the permutation procedure can be found in Appendix C. To demonstrate the applicability of the MOBIT in practice, we reanalyzed a subset of the house dust mite allergy dataset, which originally comprises pedigrees from England, Germany, Italy, and Portugal (a detailed description of the dataset can be found in Kurz et al. (2000)). The reanalyzed subset consisted of the English families, which showed a promising result with a maximum imprinting MOD score of 4.76 near the marker locus D8S511 on chromosome 8 when a model that implies complete maternal imprinting was used in the analysis (Strauch et al., 2000a). The English subset consisted of 19 families with 125 individuals, including 7 families with 2 affected sibs, 3 families with 3 affected sibs, 5 families with 4 affected sibs, and 4 extended pedigrees. One hundred fifty microsatellite markers were typed on chromosomes 1–21, with an average spacing of 10 cM at each candidate region. With regard to the sparseness of the marker

map, we used sex-specific genetic distances according to the Généthon map (Dib et al., 1996) in the analysis. As a first step, nonimprinting MOD scores and IMOD scores were calculated in a single run using the GHM software with options 'modcalc global', 'maximization dense', 'penetrance restriction off', and 'allfreq restriction off', which corresponds to a thorough evaluation of many sets of trait-model parameters in the analysis. MOD scores were calculated at all marker loci and at 9 equally spaced positions between them. Because evidence for linkage should be coupled with evidence for imprinting, we decided to interpret a significant MOBIT result to be meaningful when the corresponding IMOD score was higher than 4. For the most promising result, we obtained an empiric p value for the MOBIT using both methods *bfnm* and *perm*. With respect to method *bfnm*, replicates under the null hypothesis of linkage, but no imprinting, were generated, such that the best-fitting nonimprinting model was used for the simulation of genotype data for 1,000 replicates using SLINK (Ott, 1989; Schäffer et al., 2011; Weeks et al., 1990) and SUP (Lemire, 2006; Schäffer et al., 2011). As for method *perm*, we generated 1,000 replicates using the newly developed permutation procedure. The corresponding empiric p value for the real dataset was calculated according to the formula $p = \frac{k}{n}$, n being the total number of replicates and k the number of ordered replicates showing a MOBIT that was higher or equal to the one obtained from the real dataset.

3 Results

3.1 $H_{0,a}$: No linkage, no imprinting

The empiric 95% and 99% quantiles as well as the corresponding type I error rates assuming a χ^2 distribution with 1 df of the simulations under $H_{0,a}$: no linkage and no imprinting can be found in Table 2 (ASPs) and Table 3 (3-G pedigrees). Multipoint scenarios often showed slightly higher quantiles due to an increased effective number of tests compared to twopoint scenarios. It is of note that MOBIT quantiles for ASPs under the null hypothesis of no linkage and no imprinting are expected to show lower quantiles than the assumed χ^2 distribution with 1 df (Knapp & Strauch, 2004), which is the distribution of the MOBIT under $H_{0,b}$: linkage, but no imprinting. Due to the absence of linkage, the results did not differ between the additive and recessive trait model. The results were similar for the investigated two sample sizes. If not stated otherwise, conclusions drawn from empiric quantiles also apply to the corresponding type I error rates.

Table 2: Empiric quantiles (95% and 99%) and type I error rates (nominal $\alpha = 0.05; 0.01$) of the simulated ASP scenarios under $H_{0,a}$: No linkage, no imprinting.

													ASPs
Trait model						**Additive**						**Recessive**	
Sample size				**400**			**600**			**400**			**600**
Map ratio		**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**
Analysis using a sex-averaged map													
One marker	95%	0.4632	0.4789	0.4711	0.4514	0.4693	0.4646	0.4615	0.449	0.443	0.4574	0.4504	0.4514
	99%	0.87	0.9115	0.8596	0.8987	0.8727	0.874	0.9155	0.8625	0.8702	0.8911	0.8937	0.8903
	5%	0.0122	0.013	0.0113	0.0128	0.0121	0.0113	0.0123	0.0109	0.0114	0.0118	0.0115	0.0118
	1%	0.0016	0.0018	0.0011	0.0017	0.0009	0.001	0.0017	0.0013	0.0015	0.0015	0.0012	0.0014
Four markers, 1 cM	95%	0.5224	0.534	0.5458	0.5462	0.5326	0.5276	0.5328	0.5398	0.5407	0.5355	0.5108	0.5111
	99%	0.9569	0.9616	0.9747	0.9575	0.9584	0.9803	0.9374	0.9757	0.982	0.9602	0.9634	0.9433
	5%	0.0156	0.0159	0.0165	0.0195	0.0162	0.0154	0.0148	0.0168	0.0174	0.0167	0.0157	0.0146
	1%	0.0015	0.0016	0.0021	0.0022	0.0022	0.0018	0.0017	0.0026	0.0019	0.0021	0.0018	0.0014
Four markers, 10 cM	95%	0.6394	0.6355	0.6346	0.6278	0.6612	0.6249	0.6615	0.6373	0.6308	0.6379	0.6302	0.6326
	99%	1.1121	1.1074	1.134	1.0771	1.1162	1.113	1.1649	1.1273	1.0802	1.0901	1.0831	1.1363
	5%	0.0258	0.0259	0.0262	0.0248	0.0259	0.0256	0.0277	0.0264	0.0245	0.0267	0.0243	0.0271
	1%	0.0033	0.0032	0.0042	0.003	0.0041	0.0034	0.0038	0.004	0.0027	0.0034	0.0027	0.0035
40 SNPs	95%	0.6688	0.6352	0.6536	0.6674	0.6539	0.6399	0.6361	0.6535	0.6457	0.6378	0.6357	0.6347
	99%	1.1135	1.1125	1.1329	1.1439	1.1532	1.1963	1.0978	1.1101	1.1287	1.0549	1.1109	1.1004
	5%	0.027	0.0273	0.0261	0.028	0.0269	0.028	0.0247	0.027	0.0267	0.0235	0.0247	0.0258
	1%	0.0034	0.0038	0.0043	0.0032	0.0033	0.0046	0.0026	0.0037	0.0043	0.0035	0.0031	0.0033

Analysis using the sex-specific map as employed for the simulation													
One marker	95%	0.4785	0.2678	0.3063	0.4546	0.3115	0.2847	0.459	0.2708	0.3129	0.456	0.3142	0.315
	99%	0.9041	0.6778	0.7162	0.8863	0.6831	0.6896	0.869	0.6309	0.6989	0.8895	0.7328	0.7305
	5%	0.0128	0.0056	0.0062	0.0118	0.0056	0.0058	0.011	0.0052	0.005	0.0125	0.0065	0.0071
	1%	0.0014	0.0008	0.0005	0.0017	0.0008	0.0008	0.0011	0.0009	0.0004	0.0015	0.0006	0.0005
Four markers, 1 cM	95%	0.526	0.524	0.5168	0.5329	0.5268	0.508	0.524	0.5307	0.5034	0.529	0.5322	0.521
	99%	0.9714	0.9141	0.9531	0.968	0.9425	0.9186	0.996	0.9753	0.956	0.9858	0.9949	1.0189
	5%	0.0167	0.0138	0.0172	0.0166	0.0162	0.0143	0.018	0.0167	0.0161	0.0161	0.0164	0.017
	1%	0.0029	0.0014	0.0012	0.0015	0.0012	0.0011	0.0025	0.0023	0.0013	0.0013	0.0022	0.0023
Four markers, 10 cM	95%	0.6305	0.6315	0.6571	0.6575	0.6509	0.6647	0.6742	0.6682	0.6409	0.6396	0.6709	0.6622
	99%	1.0919	1.1001	1.1292	1.1057	1.1789	1.1542	1.1419	1.1411	1.1412	1.1343	1.1374	1.1544
	5%	0.0242	0.0251	0.0277	0.0271	0.0252	0.0281	0.0286	0.0285	0.0256	0.0259	0.0282	0.0271
	1%	0.003	0.0027	0.0037	0.0034	0.0046	0.0039	0.0035	0.0034	0.0036	0.0033	0.0038	0.0042
40 SNPs	95%	0.6517	0.6351	0.6476	0.6564	0.66	0.6309	0.6431	0.6248	0.6632	0.6212	0.6439	0.6483
	99%	1.1255	1.1045	1.1275	1.1396	1.1107	1.1787	1.156	1.1142	1.1284	1.0623	1.0784	1.1038
	5%	0.0271	0.0248	0.0285	0.0278	0.0265	0.0267	0.0261	0.0239	0.0289	0.0242	0.0255	0.027
	1%	0.0038	0.0034	0.0041	0.0036	0.0032	0.005	0.0026	0.0029	0.0035	0.0029	0.0029	0.003

Values in cM correspond to sex-averaged inter-marker distances. 95% (99%) quantile of a χ^2 distribution with 1 df, divided by $2\log(10)$: 0.8342 (1.4407). For more details see Table 1.

Table 3: Empiric quantiles (95% and 99%) and type I error rates (nominal $\alpha = 0.05; 0.01$) of the simulated 3-G pedigree scenarios under $H_{0,a}$: No linkage, no imprinting.

													3-G pedigrees
Trait model						**additive**							**recessive**
Sample size				**45**			**65**			**45**			**65**
Map ratio		**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**
Analysis using a sex-averaged map													
One marker	95%	1.1784	1.1363	1.1458	1.2075	1.2013	1.168	1.1842	1.1924	1.1754	1.1781	1.1712	1.2015
	99%	1.8483	1.7681	1.8312	1.9187	1.8871	1.8453	1.8695	1.8176	1.8187	1.812	1.7667	1.8368
	5%	0.1144	0.1108	0.1113	0.1154	0.1192	0.1164	0.1181	0.1176	0.1138	0.1189	0.1179	0.12
	1%	0.0256	0.0231	0.0258	0.0299	0.0285	0.0249	0.0274	0.0269	0.0252	0.0251	0.0243	0.0286
Four markers, 1 cM	95%	1.0853	1.0879	1.0692	1.1269	1.1052	1.092	1.1044	1.1051	1.0927	1.1092	1.1481	1.1144
	99%	1.7545	1.7816	1.7622	1.76	1.7934	1.8055	1.7989	1.7868	1.7949	1.7637	1.7961	1.8015
	5%	0.0958	0.0951	0.0935	0.1014	0.1003	0.0975	0.0965	0.0979	0.0929	0.1001	0.1025	0.0995
	1%	0.021	0.0206	0.0194	0.022	0.0229	0.0231	0.0223	0.0222	0.0214	0.0235	0.0257	0.0238
Four markers, 10 cM	95%	1.2165	1.2294	1.2033	1.2327	1.2276	1.2369	1.2221	1.2127	1.2046	1.2183	1.228	1.2448
	99%	1.8925	1.88	1.813	1.8845	1.9003	1.9725	1.8617	1.8686	1.9218	1.9072	1.8998	1.8774
	5%	0.1249	0.1212	0.1188	0.1227	0.1248	0.1293	0.1244	0.1188	0.1212	0.1256	0.1282	0.1251
	1%	0.0316	0.0318	0.0279	0.0301	0.0296	0.0323	0.0286	0.029	0.0285	0.028	0.0295	0.0291
40 SNPs	95%	1.2651	1.227	1.2115	1.232	1.2571	1.2922	1.2072	1.2255	1.2246	1.293	1.2674	1.2696
	99%	1.9924	1.9454	1.9114	1.8822	1.9504	2.0957	1.9077	1.9138	1.9407	1.9018	1.9718	1.8923
	5%	0.1311	0.1183	0.1176	0.1263	0.1312	0.1323	0.1199	0.1222	0.1217	0.1282	0.132	0.12
	1%	0.0336	0.0301	0.0293	0.0307	0.0316	0.0375	0.0291	0.0302	0.0296	0.035	0.0332	0.0307
Analysis using the sex-specific map as employed for the simulation													
One marker	95%	1.1858	1.1558	1.1208	1.1717	1.1595	1.1581	1.2106	1.1611	1.1381	1.1858	1.1819	1.1437
	99%	1.8778	1.7971	1.7768	1.8822	1.8578	1.815	1.8178	1.8463	1.7848	1.8704	1.8477	1.7957
	5%	0.1203	0.1104	0.1036	0.1112	0.1091	0.1126	0.1197	0.11	0.1064	0.1169	0.1151	0.1078
	1%	0.029	0.0254	0.0233	0.0268	0.0267	0.0259	0.0292	0.0255	0.0227	0.028	0.0254	0.025

Four markers, 1 cM	95%	1.072	1.0955	1.0979	1.1109	1.1056	1.104	1.1046	1.091	1.1037	1.1857	1.1169	1.1427
	99%	1.6991	1.8747	1.7806	1.7489	1.7816	1.816	1.7859	1.7652	1.8067	1.8625	1.7911	1.7291
	5%	0.0932	0.0979	0.0951	0.0972	0.0967	0.0982	0.0943	0.0942	0.093	0.1076	0.0979	0.0981
	1%	0.0188	0.0231	0.023	0.0227	0.024	0.0222	0.0236	0.0201	0.0223	0.272	0.0235	0.0218
Four markers, 10 cM	95%	1.2274	1.2195	1.1987	1.2344	1.2532	1.2283	1.2313	1.2519	1.225	1.2216	1.2248	1.2254
	99%	1.8604	1.989	1.8855	1.8541	1.9145	1.9445	1.9399	1.8588	1.9096	1.8539	1.8985	1.8966
	5%	0.1324	0.1244	0.1179	0.1244	0.1284	0.1313	0.1278	0.1272	0.1228	0.1299	0.1273	0.1228
	1%	0.0299	0.0304	0.0283	0.0301	0.0323	0.0327	0.0289	0.0311	0.0294	0.0289	0.0283	0.0319
40 SNPs	95%	1.2545	1.2524	1.2274	1.2462	1.2599	1.2152	1.2196	1.2122	1.2577	1.2337	1.2415	1.2623
	99%	1.9474	1.9433	1.9133	2.0043	2.0074	1.9261	1.9268	1.9235	2.0406	1.8943	1.8754	1.9926
	5%	0.1321	0.1328	0.1246	0.1302	0.1258	0.1251	0.1229	0.1218	0.1257	0.1274	0.125	0.1279
	1%	0.0327	0.0316	0.0295	0.0333	0.0342	0.0297	0.0297	0.03	0.0317	0.0301	0.0308	0.0346

Values in cM correspond to sex-averaged inter-marker distances. 95% (99%) quantile of a χ^2 distribution with 1 df, divided by $2\log(10)$: 0.8342 (1.4407). For more details see Table 1.

3.1.1 Twopoint analysis (1 marker)

With regard to ASPs, the results of the analyses using a sex-averaged map showed a $H_{0,a}$ distribution of the MOBIT with smaller quantiles than the assumed χ^2_1 distribution and showed no differences as to the underlying map ratios (Table 2). For the analyses using the sex-specific map and ASPs, empiric quantiles dropped for map ratios >1, which is due to maximization curves along the given map ratio reaching into the left half-tetrahedron $\mathbf{T}_{z_1^{pat}>z_1^{mat}}$ for ASPs in the nonimprinting likelihood (see also Methods Section 2.3 and Figure 2).

In the case of 3-G pedigrees (Table 3), the MOBIT distributions showed consistently larger quantiles than the assumed χ^2_1 distribution, irrespective of the underlying map ratio and whether a sex-averaged or a sex-specific map was used for the analysis. Obviously, the effect of maximization curves in a sex-specific MOBIT analysis using 3-G pedigrees is restricted, possibly due to peculiarities of the parameter space around the true point in terms of allele-sharing under $H_{0,a}$.

3.1.2 Multipoint analysis

The $H_{0,a}$ distribution of the MOBIT for ASPs in the two four-marker scenarios and the 40-SNPs scenario when using a sex-averaged map also had lower quantiles than the assumed χ^2_1 distribution and did not differ between the map ratios (Table 2). The corresponding sex-specific $H_{0,a}$ quantiles of the multipoint scenarios were not affected by maximization curves reaching into the left half-tetrahedron $\mathbf{T}_{z_1^{pat}>z_1^{mat}}$ for ASPs in the nonimprinting likelihood and were comparable to the quantiles of the sex-averaged analyses. This is because the maximization curves are caught between flanking markers in multipoint analyses. In addition, the maximization is multiply restricted around the true point in terms of allele-sharing under $H_{0,a}$ (see Figure 1).

For 3-G pedigrees, similar to the twopoint scenarios, the multipoint MOBIT distributions showed larger quantiles than the assumed χ^2_1 distribution, irrespective of the underlying map ratio and whether a sex-averaged or a sex-specific map was used for the analysis (Table 3).

3.2 $H_{o,b}$: *Linkage, no imprinting*

The results of the MOBIT analyses under the null hypothesis of linkage, but no imprinting can be found in Table 4 (ASPs) and Table 5 (3-G pedigrees).

Table 4: Empiric quantiles (95% and 99%) and type I error rates (nominal α = 0.05; 0.01) of the simulated ASP scenarios under $H_{0,b}$: Linkage, no imprinting.

		ASPs	
Trait model		**additive**	**recessive**

Sample size		400			600			400			600		
Map ratio		**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**
Analysis using a sex-averaged map													
One marker, 0 cM	95%	0.8127	0.8287	0.8586	0.8434	0.8329	0.8197	0.8369	0.8356	0.8419	0.836	0.8354	0.8103
	99%	1.3966	1.4299	1.487	1.4243	1.4276	1.4043	1.3917	1.4254	1.4168	1.4321	1.4884	1.3893
	5%	0.0472	0.0489	0.0539	0.0513	0.0497	0.048	0.0505	0.0501	0.0516	0.0506	0.0505	0.0461
	1%	0.009	0.0099	0.0111	0.0094	0.0095	0.0084	0.0085	0.0095	0.0091	0.0099	0.011	0.0081
One marker, 0.5 cM	95%	0.8175	0.8455	0.8625	0.8166	0.8328	0.8433	0.8273	0.8371	0.8699	0.8298	0.8595	0.8674
	99%	1.3664	1.4581	1.4589	1.4974	1.4654	1.3944	1.4032	1.4467	1.4678	1.4964	1.5154	1.4719
	5%	0.0478	0.0511	0.0539	0.047	0.0495	0.0511	0.049	0.0506	0.056	0.0497	0.0539	0.0542
	1%	0.0078	0.0108	0.0111	0.0123	0.0111	0.0094	0.0092	0.0104	0.0108	0.0115	0.0118	0.0107
One marker, 5 cM	95%	0.8229	0.8682	**1.0221**	0.833	0.886	**1.1239**	0.8455	**1.3638**	**2.4917**	0.8271	**1.582**	**3.1716**
	99%	1.4326	1.4871	**1.7243**	1.4315	1.4745	**1.8918**	1.5111	**2.1538**	**3.5398**	1.4483	**2.4018**	**4.3856**
	5%	0.0483	0.0546	**0.079**	0.05	0.0579	**0.0896**	0.0518	**0.146**	**0.4297**	0.0489	**0.1849**	**0.5909**
	1%	0.01	0.0109	**0.0189**	0.0097	0.012	**0.0248**	0.0114	**0.0431**	**0.2116**	0.0105	**0.0651**	**0.3456**
Four markers, 0.5 cM	95%	0.8798	0.8735	0.8117	0.8393	0.8897	0.8327	0.8272	0.8543	0.8245	0.8433	0.8574	0.811
	99%	1.5429	1.4836	1.4203	1.5661	1.503	1.445	1.4432	1.4284	1.4037	1.434	1.4294	1.405
	5%	0.0553	0.0558	0.047	0.0506	0.0577	0.05	0.0493	0.0531	0.0489	0.0517	0.0533	0.0474
	1%	0.0123	0.0113	0.0097	0.0123	0.0123	0.0101	0.0102	0.01	0.0091	0.0098	0.0098	0.0094
Four markers, 5 cM	95%	0.8681	0.899	0.8564	0.8913	0.9071	**0.9268**	0.8401	**0.9478**	**1.2909**	0.8519	**1.021**	**1.4261**
	99%	1.4761	1.4818	1.5257	1.5218	1.5402	**1.64**	1.3534	**1.6368**	**2.0907**	1.4882	**1.669**	**2.2446**
	5%	0.0553	0.0606	0.0541	0.0583	0.0613	**0.0636**	0.0504	**0.0667**	**0.1248**	0.0529	**0.0754**	**0.1535**
	1%	0.0111	0.0117	0.0114	0.0126	0.0135	**0.015**	0.0083	**0.0155**	**0.0373**	0.0113	**0.0164**	**0.0494**
40 SNPs	95%	0.9035	0.9085	0.9077	0.9064	0.8602	0.8695	0.8694	0.847	0.8503	0.8431	0.8307	0.8514
	99%	1.5358	1.5356	1.5856	1.4931	1.4947	1.482	1.5095	1.4936	1.4485	1.4772	1.4455	1.4426
	5%	0.059	0.0593	0.0582	0.06	0.0536	0.0546	0.0559	0.0521	0.0523	0.0514	0.0494	0.0517
	1%	0.0124	0.0133	0.0133	0.0112	0.0114	0.011	0.0121	0.0116	0.0103	0.0112	0.0103	0.0102
Analysis using the sex-specific map as employed for the simulation													
One marker, 0 cM	95%	0.8454	0.5783	0.5969	0.8367	0.5759	0.5837	0.8374	0.5563	0.604	0.8404	0.5763	0.5944
	99%	1.4425	1.1514	1.1867	1.3571	1.1316	1.1695	1.4162	1.1266	1.2076	1.4716	1.1408	1.2211
	5%	0.0519	0.0224	0.0268	0.0505	0.0242	0.0253	0.0508	0.0232	0.0263	0.0512	0.0249	0.0252
	1%	0.0102	0.005	0.0047	0.0086	0.0046	0.005	0.0096	0.0048	0.0058	0.0108	0.004	0.0057
One marker, 0.5 cM	95%	0.8229	0.5614	0.5249	0.8257	0.5563	0.5409	0.8586	0.5061	0.4335	0.8357	0.4973	0.406
	99%	1.4148	1.2026	1.0503	1.4353	1.1535	1.1095	1.4302	0.9836	0.9405	1.4658	1.1325	0.9144
	5%	0.0485	0.0243	0.0191	0.0489	0.0213	0.0213	0.0541	0.0175	0.0137	0.0504	0.0187	0.0125
	1%	0.0094	0.006	0.0038	0.01	0.0045	0.0041	0.0099	0.0027	0.002	0.0109	0.0045	0.0015
One marker, 5 cM	95%	0.8351	0.3985	0.281	0.8166	0.3937	0.2394	0.8439	0.1885	0.0074	0.861	0.1061	0.009
	99%	1.4262	0.8971	0.7075	1.397	0.8739	0.6432	1.4144	0.5228	0.1073	1.4769	0.372	0.0218
	5%	0.0502	0.0128	0.0068	0.0473	0.0111	0.0053	0.0519	0.0034	<0.0001	0.0536	0.0011	<0.0001
	1%	0.0095	0.0025	0.0008	0.009	0.0013	0.0005	0.009	0.0004	<0.0001	0.0106	<0.0001	<0.0001
Four markers, 0.5 cM	95%	0.8568	0.8711	0.8288	0.8521	0.8579	0.8194	0.8352	0.8368	0.8186	0.828	0.8318	0.8478
	99%	1.4541	1.5307	1.448	1.4211	1.4845	1.4173	1.4331	1.4486	1.3695	1.4455	1.3634	1.4573
	5%	0.0537	0.0554	0.0492	0.0527	0.0531	0.0478	0.0502	0.0507	0.0483	0.0491	0.0494	0.0522
	1%	0.0107	0.0125	0.0101	0.0094	0.0111	0.0095	0.0098	0.0102	0.0087	0.0101	0.0082	0.0104
Four markers, 5 cM	95%	0.8973	0.8649	0.8253	0.917	0.88	0.8515	0.7787	0.8621	0.8418	0.8329	0.867	0.8937
	99%	1.4615	1.4827	1.3823	1.5867	1.5335	1.4475	1.4048	1.4894	1.4061	1.4754	1.5133	1.5237
	5%	0.0586	0.0555	0.0486	0.0621	0.0577	0.052	0.0421	0.0541	0.0504	0.0498	0.0543	0.0582
	1%	0.0106	0.0114	0.0086	0.0136	0.0124	0.0106	0.0087	0.0112	0.0091	0.0106	0.0114	0.0127

40 SNPs	95%	0.9108	0.8987	0.8362	0.8941	0.8774	0.8419	0.8582	0.8432	0.828	0.8312	0.8153	0.8565
	99%	1.5531	1.4804	1.4479	1.5509	1.4793	1.4533	1.4704	1.4238	1.3847	1.4183	1.3961	1.4457
	5%	0.0616	0.0599	0.0502	0.0574	0.0546	0.0515	0.0529	0.0516	0.049	0.0497	0.0471	0.0528
	1%	0.0131	0.0111	0.0102	0.0138	0.0111	0.0105	0.0113	0.0097	0.0087	0.0093	0.009	0.0102

Values in cM correspond to sex-averaged marker-trait locus distances. 95% (99%) quantile of a χ^2 distribution with 1 df, divided by $2\log(10)$: 0.8342 (1.4407). Values in **bold** indicate inflated quantiles and type I error rates due to confounding between sex-specific recombination fractions and genomic imprinting. For more details see Table 1.

Table 5: Empiric quantiles (95% and 99%) and type I error rates (nominal $\alpha = 0.05; 0.01$) of the simulated 3-G pedigree scenarios under $H_{0,b}$: Linkage, no imprinting.

													3-G pedigrees
Trait model							**additive**						**recessive**
Sample size				**45**			**65**			**45**			**65**
Map ratio		**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**	**1:1**	**7:3**	**9:1**
Analysis using a sex-averaged map													
One marker, 0 cM	95%	1.0483	1.0374	1.0672	0.9659	1.0291	0.9771	0.8145	0.8082	0.821	0.8021	0.8145	0.8275
	99%	1.8388	1.7556	1.7557	1.6499	1.7091	1.6911	1.3563	1.3998	1.4	1.3584	1.4378	1.354
	5%	0.0791	0.0802	0.0871	0.0696	0.0777	0.0741	0.0474	0.0467	0.0476	0.0447	0.0477	0.0493
	1%	0.0214	0.0201	0.0199	0.018	0.0176	0.0172	0.0076	0.0087	0.0085	0.0081	0.0099	0.0084
One marker, 0.5 cM	95%	1.0655	1.0573	1.055	0.9983	0.9878	1.0228	0.8253	0.8433	0.8362	0.8157	0.8685	0.827
	99%	1.7972	1.6659	1.7346	1.6928	1.7089	1.7136	1.3968	1.4412	1.4075	1.4293	1.4628	1.4281
	5%	0.0874	0.0855	0.0813	0.075	0.0726	0.0759	0.0493	0.0519	0.0506	0.0485	0.0549	0.0489
	1%	0.0231	0.0189	0.0208	0.0185	0.019	0.0194	0.009	0.101	0.0095	0.0099	0.0107	0.0096
One marker, 5 cM	95%	1.0941	1.1011	1.0998	1.0668	1.0603	1.1036	0.8632	**0.9687**	**1.2511**	0.9032	**1.026**	**1.4039**
	99%	1.7775	1.8277	1.7539	1.7211	1.8168	1.7925	1.4568	**1.6494**	**2.075**	1.5279	**1.7263**	**2.2133**
	5%	0.0909	0.0916	0.0953	0.0911	0.0855	0.0916	0.0535	**0.0686**	**0.1115**	0.0607	**0.0778**	**0.144**
	1%	0.0222	0.0248	0.0235	0.0222	0.0221	0.0233	0.0104	**0.0164**	**0.0347**	0.0129	**0.0194**	**0.0468**
Four markers, 0.5 cM	95%	0.9891	1.0036	0.9493	0.9584	0.9382	0.9408	0.8354	0.808	0.7827	0.8024	0.8139	0.775
	99%	1.7004	1.6667	1.6085	1.6035	1.6421	1.6464	1.444	1.3915	1.3825	1.3559	1.495	1.3931
	5%	0.0741	0.0755	0.0653	0.0668	0.0648	0.0646	0.0501	0.0462	0.0439	0.0462	0.0476	0.0419
	1%	0.0167	0.0167	0.0142	0.0147	0.0154	0.0161	0.0101	0.0092	0.0083	0.0083	0.0115	0.0089
Four markers, 5 cM	95%	1.0746	1.0712	1.0277	1.0109	1.027	1.0074	0.7978	0.8499	**0.95**	0.806	0.8633	**0.9306**
	99%	1.7492	1.7618	1.7313	1.6572	1.6982	1.6708	1.4304	1.4569	**1.5372**	1.3541	1.4902	**1.6792**
	5%	0.0848	0.0887	0.0788	0.0758	0.0781	0.0777	0.0461	0.052	**0.0653**	0.0465	0.0541	**0.0638**
	1%	0.0209	0.023	0.0182	0.018	0.0195	0.0169	0.0096	0.0106	**0.0138**	0.009	0.0116	**0.0161**
40 SNPs	95%	1.0386	1.0255	1.0074	0.9789	0.9925	0.9815	0.7949	0.8175	0.817	0.8381	0.8209	0.8072
	99%	1.7208	1.726	1.7	1.58	1.619	1.7236	1.4125	1.4452	1.4265	1.4346	1.4263	1.4496
	5%	0.0819	0.077	0.0741	0.0726	0.0715	0.0709	0.045	0.0477	0.047	0.0507	0.0482	0.0471
	1%	0.0191	0.0191	0.0184	0.0147	0.0155	0.0177	0.0092	0.0102	0.0097	0.0099	0.0095	0.0103
Analysis using the sex-specific map as employed for the simulation													
One marker, 0 cM	95%	1.0608	0.9992	0.9616	0.9594	0.9349	0.9317	0.8177	0.7607	0.7583	0.8074	0.797	0.7729
	99%	1.7583	1.6683	1.5927	1.6848	1.5877	1.6011	1.3415	1.3163	1.3395	1.3737	1.3355	1.3541
	5%	0.0792	0.0757	0.0687	0.0701	0.0644	0.0658	0.0482	0.0406	0.0391	0.0461	0.0447	0.0393
	1%	0.0192	0.0174	0.0141	0.0173	0.0136	0.0135	0.0072	0.0069	0.0076	0.0083	0.0071	0.0084
One marker, 0.5 cM	95%	1.0663	0.9906	0.9635	0.9749	0.9458	0.9498	0.8186	0.7503	0.716	0.8314	0.7356	0.7102
	99%	1.7535	1.6446	1.6307	1.6397	1.5692	1.6368	1.3805	1.2754	1.2609	1.3691	1.2924	1.3014
	5%	0.0817	0.0723	0.0711	0.0712	0.0667	0.0663	0.0473	0.0378	0.0359	0.0496	0.0364	0.0358
	1%	0.0214	0.0171	0.0156	0.0176	0.0149	0.0154	0.0089	0.0067	0.0055	0.0082	0.0063	0.007
One marker, 5 cM	95%	1.1129	0.9857	0.974	1.0475	0.9476	0.9565	0.8493	0.719	0.658	0.826	0.7171	0.6731

	99%	1.7828	1.6597	1.6038	1.741	1.6496	1.6168	1.4627	1.2505	1.167	1.4412	1.17	1.1188
	5%	0.0941	0.0732	0.0731	0.0855	0.0679	0.0689	0.0526	0.035	0.0288	0.0487	0.0325	0.0299
	1%	0.0232	0.0173	0.0158	0.0189	0.017	0.0157	0.0106	0.0065	0.0036	0.0101	0.0055	0.0029
Four markers, 0.5 cM	95%	0.9874	1.0233	0.9818	0.9555	0.9942	0.938	0.8088	0.8155	0.8084	0.7908	0.7968	0.8094
	99%	1.7174	1.7472	1.6581	1.6428	1.7033	1.6004	1.3616	1.3805	1.4248	1.3701	1.3075	1.4083
	5%	0.0723	0.0782	0.0723	0.0682	0.0717	0.0651	0.0474	0.0469	0.0466	0.0436	0.0456	0.0463
	1%	0.0172	0.0186	0.0162	0.0161	0.0184	0.0144	0.0077	0.0086	0.0096	0.0086	0.0075	0.0095
Four markers, 5 cM	95%	1.0651	1.0691	1.0344	1.0168	1.0063	0.9618	0.8206	0.825	0.7817	0.8199	0.822	0.7875
	99%	1.7923	1.7826	1.7352	1.7108	1.61	1.6321	1.4158	1.4263	1.4319	1.4207	1.427	1.3247
	5%	0.0872	0.0878	0.0804	0.0764	0.0755	0.0682	0.0486	0.0482	0.0435	0.0484	0.0481	0.0424
	1%	0.0221	0.0205	0.0193	0.0188	0.016	0.0168	0.0094	0.0097	0.0099	0.0098	0.0098	0.0072
40 SNPs	95%	1.0363	0.9697	1.0075	0.9758	1.0128	0.9701	0.8032	0.8068	0.8105	0.8163	0.8373	0.8135
	99%	1.6783	1.6172	1.6526	1.665	1.6832	1.563	1.3855	1.3863	1.4469	1.3599	1.4731	1.4228
	5%	0.0831	0.0718	0.0746	0.0699	0.0765	0.0676	0.0468	0.0466	0.0471	0.0473	0.0508	0.0473
	1%	0.0185	0.0162	0.0176	0.0162	0.0184	0.0141	0.0086	0.0091	0.0102	0.008	0.011	0.0089

Values in cM correspond to sex-averaged marker-trait locus distances. 95% (99%) quantile of a χ^2 distribution with 1 df, divided by $2\log(10)$: 0.8342 (1.4407). Values in **bold** indicate inflated quantiles and type I error rates due to confounding between sex-specific recombination fractions and genomic imprinting. For more details see Table 1.

3.2.1 Twopoint, sex-averaged analysis

With respect to ASPs and the additive as well as the recessive trait model, the $H_{0,b}$ distribution of the MOBIT for the scenario with 0 cM between marker and trait locus corresponded well to the expected χ^2 distribution with 1 df for all map ratios and using a sex-averaged map in the analysis (Table 4). Although the additive trait model lies on the boundary of the tetrahedron **T** in terms of allele-sharing for ASPs, it is readily conceivable that the two front planes in Figure 1 (left side) are explored in two dimensions with the IMOD score and in one dimension with the MOD score, corresponding to a difference in maximized parameters of $2 - 1 = 1$, which is the same as for the interior of **T** ($3 - 2 = 1$). Taken together, this results in empiric MOBIT quantiles equal to those of a χ^2_1 distribution. The corresponding quantiles for the scenario with 0.5 cM between marker and trait locus for both trait models corresponded to the expected χ^2_1 distribution for the 1:1 map ratio, with some inflation when moving to the 7:3 and further to the 9:1 map ratio due to confounding. In the case of the scenario with 5 cM between marker and trait locus, MOBIT quantiles were slightly inflated for the 7:3 map ratio and clearly inflated for the 9:1 map ratio due to confounding (Table 4). Confounding was more severe for the larger sample size, the recessive trait model, and the larger marker-trait locus distance.

In the case of 3-G pedigrees and the additive trait model, the MOBIT $H_{0,b}$ quantiles of the twopoint scenarios were clearly inflated compared to the expected χ^2 distribution with 1 df (Table 5). This is probably because the true point in terms of allele-sharing of the additive model also lies on the boundary of the parameter space of 3-G pedigrees as it is for ASPs. In contrast to ASPs, however, this apparently does not lead to 1 df for the MOBIT, presumably because the parameter space of 3-G pedigrees has a more complicated form with more boundaries. Further, the inflation increased with increasing marker-trait locus distance, however, additional inflation of quantiles due to confounding could not be observed (Table 5). With regard to the recessive model, the $H_{0,b}$ distribution of the MOBIT for the scenarios with 0 and 0.5 cM between marker and trait locus corresponded well to the expected χ^2 distribution with 1 df for all map ratios and using a sex-averaged map in the analysis (Table 5). For the scenario with 5 cM between marker and trait locus, however, inflation of MOBIT quantiles and hence increased type I error rates due to confounding could be observed.

As can be seen from Table 4 and Table 5, confounding of twopoint scenarios was more severe for ASPs than for 3-G pedigrees, especially for the recessive trait model (observed type I error rates for the 5 cM twopoint scenario and a 9:1 map ratio assuming a χ^2_1 distribution with a nominal $\alpha = 5\%$ significance level: 0.5909 and 0.144 for 600 ASPs and 65 3-G pedigrees, respectively).

3.2.2 Twopoint, sex-specific analysis

With regard to ASPs and the scenario with 0 cM between marker and trait locus, MOBIT quantiles were significantly lower than those for the expected χ^2_1 distribution due to maximization curves reaching into the left

half-tetrahedron $\mathbf{T}_{z_i^{pat}>z_i^{mat}}$ for ASPs in the nonimprinting likelihood for map ratios >1 (Table 4). Quantiles were even lower for the scenario with 0.5 cM between marker and trait locus and map ratios >1, especially for the recessive trait model. The results for the 0.5 cM scenario were comparable between the two sample sizes. In the case of the scenario with 5 cM between marker and trait locus, the deflation of quantiles was increasingly severe for the larger map ratio, the larger sample size, and the recessive trait model.

With 3-G pedigrees, MOBIT quantiles were deflated with an increasing map ratio and an increasing marker-trait locus distance for both trait models (Table 5). The degree to which maximization curves according to a given sex-specific map ratio deflate the MOBIT was smaller for 3-G pedigrees compared to ASPs. This might be due to the complexity of the parameter space for 3-G pedigrees, such that more spatial restrictions prevent the maximization curves from reaching farther into the interior of the parameter space.

3.2.3 Multipoint, sex-averaged analysis

In the case of the 4 markers, 0.5 cM scenario, the $H_{0,b}$ distribution of the MOBIT for all map ratios and both trait models corresponded to the expected χ^2 distribution with 1 df for ASPs (Table 4). As explained above, slightly higher quantiles for multipoint analysis are due to the increased effective number of tests compared to twopoint analyses. With respect to the 4 markers, 5 cM scenario and the additive trait model, quantiles were slightly inflated due to confounding for the larger sample size and a map ratio of 9:1. In the case of the recessive trait model, quantiles were increasingly inflated due to confounding for larger map ratios and sample sizes. Again, confounding was more severe for the recessive trait model. The scenario with 40 SNPs and a marker spacing of 0.32 cM did not show confounding.

In summary, a marker spacing of 1 cM (corresponding to a marker-trait locus distance of 0.5 cM to both flanking markers) seemed to be sufficient to avoid confounding, even in the case of such an extreme map ratio as 9:1.

In the case of 3-G pedigrees and the additive trait model, MOBIT quantiles were higher than compared to the expected χ^2 distribution with 1 df for all multipoint scenarios (Table 5) due to the above mentioned boundary conditions of the true point in terms of allele sharing in the parameter space of 3-G pedigrees. Slightly higher MOBIT quantiles were obtained for the 4 markers, 5 cM scenario compared to the 4 markers, 0.5 cM scenario. The quantiles of the 40 SNPs scenario mostly lay between the two other multipoint scenarios. Despite the different quantiles for the three marker settings, there was no evidence for confounding for all multipoint scenarios and the additive trait model. The corresponding quantiles for the recessive trait model corresponded to the expected χ^2 distribution with 1 df for all map ratios, except for the 4 markers, 5 cM scenario and a map ratio of 9:1 due to confounding (Table 5).

Again, confounding was more severe for ASPs than for 3-G pedigrees, especially for the recessive trait model (type I error rates for the 4 markers, 5 cM scenario and a 9:1 map ratio assuming a χ^2_1 distribution with a nominal $\alpha = 5\%$ significance level: 0.1535 and 0.0638 for 600 ASPs and 65 3-G pedigrees, respectively).

3.2.4 Multipoint, sex-specific analysis

Due to the multipoint setting, in which the putative trait locus is confined between flanking markers, such that the outreach of maximization curves is limited, all investigated multipoint distributions roughly corresponded to the expected χ^2 distribution with 1 df for ASPs (Table 4). There was only a slight deflation of quantiles for the multipoint scenarios with a 9:1 map ratio and the additive trait model. The results did not substantially differ between the two sample sizes.

In the case of the 3-G pedigrees and the additive trait model, MOBIT quantiles were higher than expected as was explained above. Quantiles only slightly decreased with increasing map ratio (Table 5). In the case of the recessive model, the distributions roughly corresponded to the expected χ^2 distribution with 1 df. Apart from the 4 markers, 5 cM scenario, MOBIT quantiles were not deflated due to a map ratio >1.

3.3 *H_1: Linkage, imprinting*

The results of the power calculations for the twopoint scenarios and the 40 SNPs, 0.32 cM scenario can be found in Figure 4 and Figure 5. The critical values for a test with a true type I error rate of 5% corresponded to the respective $H_{0,b}$ 95% quantiles of each particular scenario (Table 4 and Table 5 for ASPs and 3-G pedigrees, respectively). In general, power was higher for the larger sample size for both pedigrees.

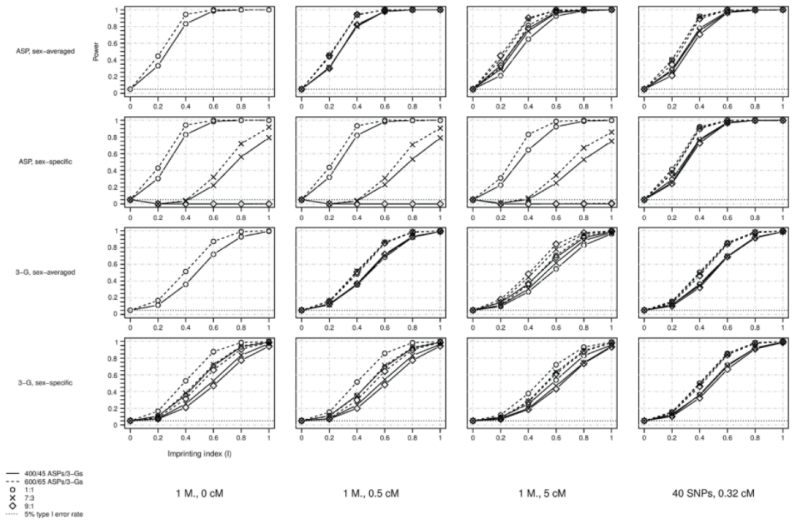


Figure 4: Power to detect imprinting using the MOBIT. For more details see Figure 3.

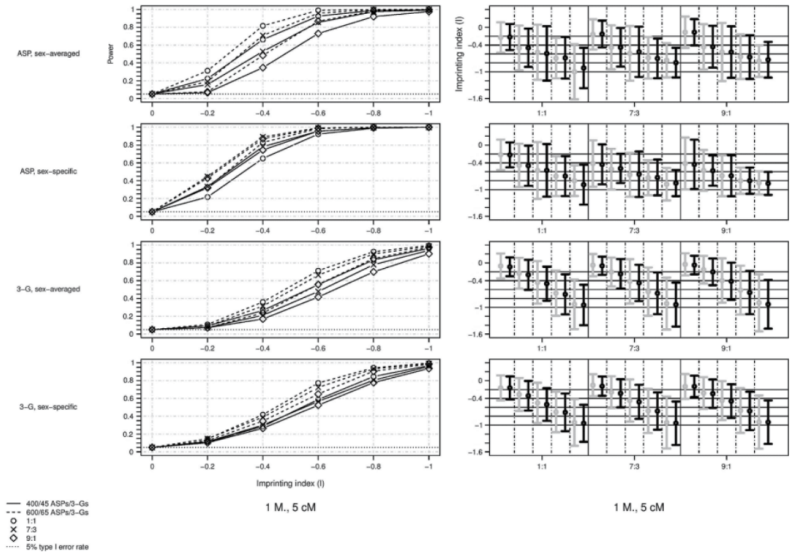


Figure 5: Power and estimation accuracy of imprinting index I of the MOBIT for the paternal imprinting model. Median estimates of I are depicted as bullets, the corresponding median absolute deviation (MAD; adjusted by a constant (1.4826) for asymptotically normal consistency) is shown as error bars, and the larger sample size is colored in black. The scenarios are sorted in increasing order based on the absolute value of I from left to right. For more details see Figure 3.

3.3.1 Twopoint, sex-averaged analysis, maternal imprinting

The power results for the maternal imprinting model using a sex-averaged map in the analysis can be found in Figure 4 (columns 1–3, rows 1 and 3 for ASPs and 3-G pedigrees, respectively). For the 1 marker, 0 cM scenario, results for map ratios larger than 1 are not shown, because if $\theta = 0$ between marker and trait locus, the existence of sex-specific recombination fractions in the marker-surrounding genetic region is of no relevance when using a sex-averaged map for the analysis (see also Table 4 and Table 5 for the respective type I error rates under $H_{0,b}$).

The power of the MOBIT was not affected by the underlying map ratio in the case of the 0.5 cM scenario across both pedigree types. However, for the 5 cM scenario, larger map ratios showed slightly higher power than smaller ones across both pedigree types. In general, power was higher for ASPs compared to 3-G pedigrees across all investigated twopoint scenarios and was slightly lower for larger marker-trait locus distances. In the case of the 0 cM and 0.5 cM scenarios, a power consistently >80% was obtained for an imprinting index $I = 0.4$ with ASPs for both sample sizes, whereas power was consistently >80% for $I = 0.8$ with 3-G pedigrees for both sample sizes. The corresponding values to obtain a power consistently >80% for the 5 cM scenario were $I = 0.6$ with ASPs for both sample sizes and $I = 0.8$ with 3-G pedigrees for both sample sizes.

3.3.2 Twopoint, sex-averaged analysis, paternal imprinting

The results of the power analysis for the paternal imprinting model and a marker-trait locus distance of 5 cM can be found in Figure 5 (left). For small to moderate imprinting degrees and a map ratio larger than 1, power was lower compared to the corresponding maternal imprinting scenarios in Figure 4 for both pedigree types. This is due to the fact that, in the case of ASPs, the true point in terms of allele-sharing is gradually shifted from $\mathbf{T}_{z_1^{pat} > z_1^{mat}}$ into $\mathbf{T}_{z_1^{pat} < z_1^{mat}}$ with increasing imprinting degrees, thereby crossing the possible triangle. In the case of the 1:1 map ratio, however, the corresponding points in terms of allele-sharing instantly lie within $\mathbf{T}_{z_1^{pat} < z_1^{mat}}$ even for small imprinting degress and can exclusively be reached by the IMOD score maximization. A similar behaviour was observed for 3-G pedigrees.

3.3.3 Multipoint, sex-averaged analysis, maternal imprinting

The power results of the MOBIT for the multipoint scenarios differed only slightly between the map ratios and were generally higher for larger sample sizes (Figure 4, column 4, rows 1 and 3 for ASPs and 3-G pedigrees, respectively). Power was higher for ASPs compared to 3-G pedigrees. More specifically, a power consistently >80% was obtained for an imprinting index $I = 0.6$ with ASPs for both sample sizes, whereas power was consistently >80% for $I = 0.8$ with 3-G pedigrees for both sample sizes.

3.3.4 Twopoint, sex-specific analysis, maternal imprinting

Regarding the results of the MOBIT for ASPs and the maternal imprinting model, the scenarios with a map ratio of 1:1 had the highest power, with higher values for larger sample sizes and higher imprinting degrees, followed by the scenarios with a map ratio of 7:3 (Figure 4, columns 1–3, row 2). The twopoint scenarios with a map ratio of 9:1 had no power to detect imprinting, which was due to the problem of maximization curves (Figure 2). The power even dropped below the nominal type I error rate of 5%. This can be explained by the fact that under $H_{0,\,b}$, where $I = 0$, a MOBIT greater zero results mainly from maxima in terms of allele sharing due to sampling variation of the simulated replicates in the right half-tetrahedron $\mathbf{T}_{z_1^{pat} < z_1^{mat}}$ (for ASPs), which is exclusively explored by the imprinting (IMOD score) analysis. An imprinting effect $I > 0$ shifts the point into the left half-tetrahedron $\mathbf{T}_{z_1^{pat} > z_1^{mat}}$, leading to maxima in the right half-tetrahedron being less likely. Due to the maximization curves, with an assumed map ratio of 9:1, the sample maxima are covered by the nonimprinting (MOD score) maximization as well as by the imprinting maximization, resulting in a power below 5%. In the case of the 7:3 map ratio, the same effect is appreciable for $I = 0.2$, whereas for $I \geq 0.4$, the stronger imprinting outweighs this effect of maximization curves. In the case of ASPs, there was almost no difference in power between the scenarios with varying marker-trait locus distances (0 cM, 0.5 cM, and 5 cM) for map ratios >1.

With 3-G pedigrees, the problem of maximization curves seemed to be smaller. The scenarios with a 1:1 map ratio and a larger sample size still had the highest power (Figure 4, columns 1–3, row 4). Interestingly, the other two map ratios showed comparable power, with again slightly higher power for scenarios with a map ratio of 7:3 compared to 9:1. In the case of 3-G pedigrees, power was slightly lower for larger marker-trait locus distances. In addition, power was consistently higher for 3-G pedigrees compared to ASP analyses in the case of a map ratio >1.

3.3.5 Twopoint, sex-specific analysis, paternal imprinting

The power for the paternal imprinting model when using a sex-specific map in the analysis (see Figure 5, left) was substantially higher compared to the respective maternal imprinting model for both pedigree types as depicted in Figure 4. This is because maximization curves of the nonimprinting MOD score are restricted to points in terms of allele-sharing that correspond to excess paternal sharing, such that points corresponding to excess maternal sharing are exclusively covered by the IMOD score maximization. Further, the power was higher with increasing map ratio for ASPs due to smaller empiric threshold values for the MOBIT as derived from the respective $H_{0,\,b}$ simulations (see Table 4). The power was also higher compared to the corresponding sex-averaged analyses, for which empiric threshold values for the MOBIT under $H_{0,\,b}$ were inflated due to confounding (see Table 4). Conversely, the power of the 3-G pedigress was comparable to that of the respective sex-averaged analyses and was slightly higher for smaller map ratios, similar to the findings in Results Section 3.3.4.

3.3.6 Multipoint, sex-specific analysis, maternal imprinting

The power results of the MOBIT for the multipoint scenarios depended only slightly on the map ratio, because the putative disease locus is confined between flanking markers, which largely avoids maximization curves. Power values for the 1:1 map ratio were somewhere between the 1 marker, 0.5 cM and 1 marker, 5 cM scenarios with higher power observed for larger sample sizes and ASPs (Figure 4, column 4, rows 2 and 4).

3.4 *Estimation of imprinting index I in a sex-averaged MOBIT analysis*

3.4.1 Twopoint analysis, maternal imprinting

The twopoint imprinting parameter estimation results of the sex-averaged MOBIT analyses for ASPs can be found in Figure 6 (rows 1–3). The estimated median imprinting indices were close to their expected values for the 0 and 0.5 cM scenarios. In the case of the 5 cM scenario, imprinting indices <0.6 were mostly overestimated, whereas imprinting indices >0.6 were mostly underestimated. For the larger sample size, the underestimation was less pronounced. For a given map ratio, the variation as measured by the median absolute deviation (*MAD*) was highest for $I = 0.6$ and lowest for $I = 0.2$. Further, *MAD* was slightly lower for the larger sample size for most investigated scenarios.

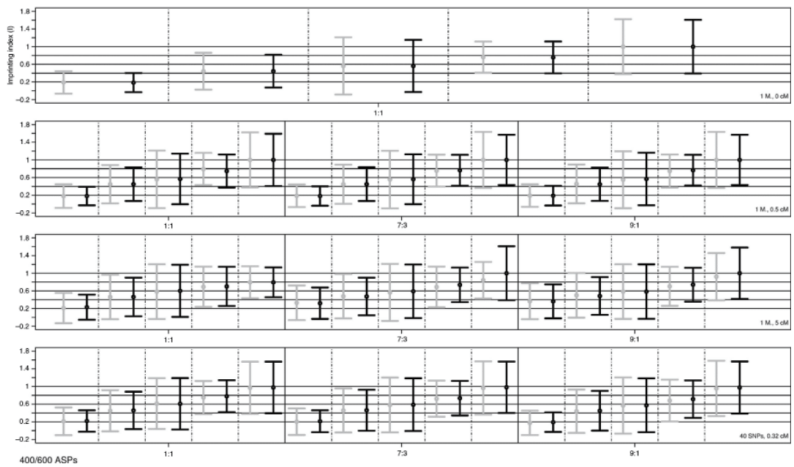


Figure 6: Estimation of imprinting index *I* by the MOBIT using a sex-averaged map in the analysis of ASPs. For more details see Figure 3 and Figure 5.

The corresponding imprinting parameter estimation results for 3-G pedigrees can be found in Figure 7 (rows 1–3). Median values of the estimated imprinting indices were close to their expected values, although

underestimated, especially in the case of lower imprinting indices. *MAD* was lowest for $I = 0.2$ and slightly increased with larger imprinting indices and larger marker-trait locus distances, but did not substantially differ between map ratios of a given marker-trait locus distance. *MAD* was lower for the larger sample size for all investigated scenarios. For most scenarios, *MAD* of the imprinting index was lower for 3-G pedigrees compared to ASPs.

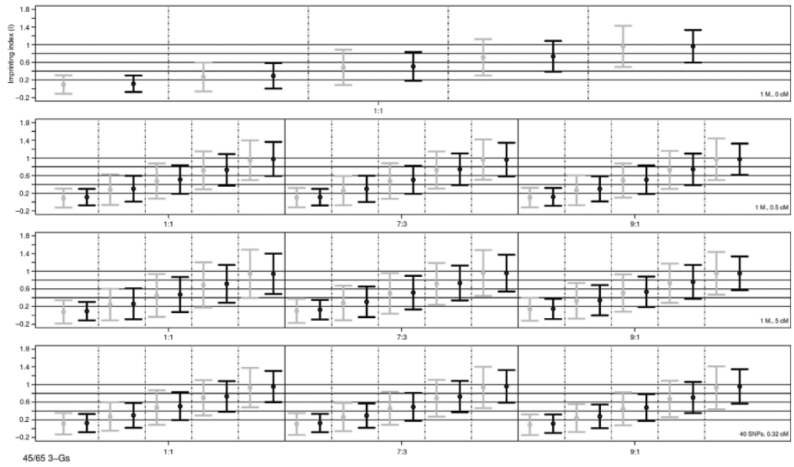


Figure 7: Estimation of imprinting index *I* by the MOBIT using a sex-averaged map in the analysis of 3-G pedigrees. For more details see Figure 3 and Figure 5.

3.4.2 Twopoint analysis, paternal imprinting

The parameter estimation results for the paternal imprinting model using a sex-averaged map in the analysis can be found in Figure 5 (right, rows 1 and 3). The results for both pedigree types and the 1:1 map ratio were comparable to those of the respective maternal imprinting scenario (see Figure 6 and Figure 7, row 3). Specifically, the point estimate of the imprinting index was either close to or slightly less negative than its expected value. In the case of a map ratio larger than 1, estimates were often less negative than the expected values. This is because excess maternal allele-sharing at the disease locus is attenuated by a longer female genetic map, which reduces the maternal sharing excess at the marker locus.

3.4.3 Multipoint analysis

The multipoint results of the imprinting parameter estimation for the sex-averaged MOBIT analyses using ASPs can be found in Figure 6 (row 4). Estimated median imprinting indices were close to their expected values across all map ratios. *MAD* was highest for $I = 0.6$ and lowest for $I = 0.2$. *MAD* was lower for the larger sample size, for most investigated scenarios, and did not substantially differ between map ratios.

The corresponding results for 3-G pedigrees are shown in Figure 7 (row 4). Imprinting indices were slightly underestimated across all map ratios. The underestimation was even a bit stronger than for the twopoint scenarios in the case of the 7:3 and 9:1 map ratios. *MAD* was lowest for $I = 0.2$ and slightly increased with larger imprinting indices. Again, *MAD* was lower for the larger sample size and was comparable with respect to different map ratios. In addition, *MAD* of the imprinting index was lower for 3-G pedigrees compared to ASPs.

3.5 *Estimation of imprinting index I in a sex-specific MOBIT analysis*

3.5.1 Twopoint analysis, maternal imprinting

The twopoint results of the estimation of the imprinting index *I* using a sex-specific map in a MOBIT analysis using ASPs can be found in Figure 8 (rows 1–3). Apart from differences due to sampling variation, the results

for the 1:1 map ratio were identical to the corresponding scenarios in Figure 6 for the sex-averaged analysis (column 1, rows 1–3). With increasing map ratio, however, imprinting indices were significantly underestimated, especially in the case of smaller imprinting degrees ($I < 0.8$). This was due to maximization curves reaching into the left half-tetrahedron $\mathbf{T}_{z_1^{pat} > z_1^{mat}}$, as it was explained above (Results Section 3.3.4), which also applies to the maximization under imprinting. In the case of a map ratio >1, the corresponding variation in terms of *MAD* was slightly higher for larger imprinting indices, compared to the corresponding sex-averaged MOBIT analysis (Figure 6, rows 1–3), but markedly lower for the scenarios with smaller *I*, in which imprinting indices were significantly underestimated. *MAD* was comparable between the two sample sizes.

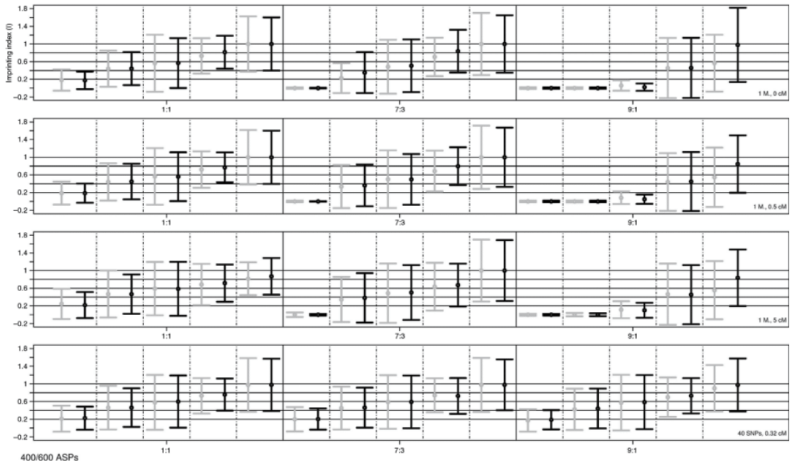


Figure 8: Estimation of imprinting index *I* by the MOBIT using a sex-specific map in the analysis of ASPs. For more details see Figure 3 and Figure 5.

The results of the imprinting parameter estimation using 3-G pedigrees are depicted in Figure 9 (rows 1–3). As with ASPs, results for the 1:1 map ratio with 3-G pedigrees were identical to the corresponding scenarios in Figure 7 for the sex-averaged analysis (column 1, rows 1–3). Estimated median imprinting indices were underestimated across all map ratios and marker-trait locus distances. *MAD* increased from $I = 0.2$ over $I = 0.4$ to $I = 0.6$, with comparable variation for $I \geq 0.6$. Further, *MAD* slightly increased with larger marker-trait locus distance and map ratio. Compared to the results for ASPs (Figure 8, rows 1–3), a better imprinting parameter estimation accuracy in terms of lower bias and variability was obtained for 3-G pedigrees and map ratios >1. This was in accordance with the finding that 3-G pedigrees showed higher power than ASPs in a sex-specific MOBIT analysis for map ratios >1, probably due to a reduced effect of maximization curves (see Results Section 3.3.4).

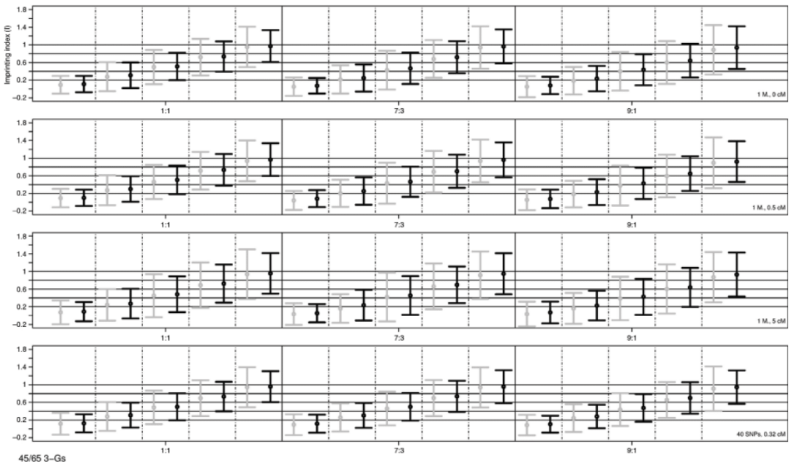


Figure 9: Estimation of imprinting index I by the MOBIT using a sex-specific map in the analysis of 3-G pedigrees. For more details see Figure 3 and Figure 5.

3.5.2 Twopoint analysis, paternal imprinting

The imprinting index estimation results for the paternal imprinting model using a sex-specific map in the analysis can be found in Figure 5 (right, rows 2 and 4) for both pedigree types. Estimation accuracy was generally better compared to the maternal imprinting scenario, especially for ASPs (see also Figure 8). This is due to a less distinguished effect of maximization curves, which particularly affects allele-sharing points in the other half tetrahedron $\mathbf{T}_{z_1^{pat}>z_1^{mat}}$. In the case of ASPs, imprinting indices for scenarios with $I < 0.6$ were overestimated, whereas those with $I > 0.6$ were underestimated. The corresponding results for 3-G pedigrees, however, were similar to the sex-averaged analysis due a less pronounced effect of maximization curves compared to ASPs.

3.5.3 Multipoint analysis

The results of the multipoint scenarios using a sex-specific map in the analysis for ASPs can be found in Figure 8 (row 4). The results were similar to the corresponding sex-averaged analyses (Figure 6, row 4). In particular, estimated median imprinting indices were close to their expected values across all map ratios. *MAD* was highest for $I = 0.6$ and lowest for $I = 0.2$. *MAD* was lower for the larger sample size for most scenarios and did not substantially differ between map ratios.

The multipoint results for 3-G pedigrees are shown in Figure 9 (row 4). As with ASPs, the results were similar to the corresponding sex-averaged analyses (Figure 7, row 4). In particular, imprinting indices were slightly underestimated across all map ratios. *MAD* was lowest for $I = 0.2$ and slightly increased with larger imprinting indices. Again, *MAD* was lower for the larger sample size and was comparable with respect to different map ratios. In addition, *MAD* of the imprinting index was lower for 3-G pedigrees compared to ASPs (Figure 8, row 4).

3.6 *Real data example on house dust mite allergy*

The MOBIT results for all investigated chromosomes can be found in Table 6. Assuming an IMOD score larger than 4 to be an informative linkage signal, an empiric p value was simulated using the two proposed MOBIT simulation/permutation procedures for the result on chromosome 8 near marker D8S511 (IMOD score = 4.9675, MOD score = 3.186, MOBIT = 1.7815, best-fitting nonimprinting model: $p = 0.07$, $f_0 = 0.0$, $f_1 = 0.0001$, $f_2 = 0.0$; best-fitting imprinting model: $p = 0.14$, $f_0 = 0.002$, $f_{1,pat} = 0.06$, $f_{1,mat} = 0.0$, $f_2 = 0.0$, maternal imprinting). The empiric p value obtained using method *bfnm* was p = 0.02, whereas it was p = 0.18 using method *perm*. We also repeated the analysis using a sex-averaged map and obtained similar results (IMOD score = 5.1301,

MOD score = 3.2081, MOBIT = 1.9219, p = 0.005 using method *bfnm* and p = 0.178 using method *perm*). This finding illustrates that there may exist various ways to construct a valid simulation/permutation test, whereby the choice that maximizes power with simultaneous control over the type I error rate presumably depends on the underlying null hypothesis, the actual trait model and investigated pedigree types (see also Churchill and Doerges (2008)). In the case of the house dust mite data, method *perm* might have been less powerful than method *bfnm*. The corresponding results for the investigated scenarios in Appendix C also showed differences between the two methods for both sex-averaged and sex-specific analyses. In the case of ASPs and when the correct sex-specific map was used in the analysis, differences might be due to the fact that the conditional null hypothesis underlying the *perm* method implies an invariant nonimprinting maximum likelihood estimate for the point in terms of allele-sharing ('preserved linkage'), such that there is no effect of maximization curves for the permuted replicates in a sex-specific nonimprinting MOD score analysis. Therefore, permuted replicates almost always lead to points in terms of allele-sharing that are exclusively reached by the IMOD score imprinting maximization (see Appendix Figure 11). As to the method *bfnm*, due to maximization curves, the nonimprinting maximum likelihood estimate has good chance to be close to the one obtained by the imprinting maximization for every simulated replicate, which leads to smaller quantiles of the null distribution compared to method *perm*.

Table 6: Results of the MOBIT real data application on house dust mite allergy using a sex-specific map in the analysis.

Chromosome	Length in cM	MOD	IMOD	MOBIT
1	238.3	2.5933	2.5933	0
2	240.8	1.9577	3.1872	1.2294
3	147.7	0.8161	2.2488	1.4327
4	153.6	3.7896	3.7991	0.0095
5	140.3	3.1719	3.1771	0.0051
6	154.4	2.4707	3.2292	0.7584
7	124.9	2.772	3.2912	0.5193
8	**41.7**	**3.186**	**4.9675**	**1.7815**
9	0	0.602	0.602	0
10	129.3	2.2501	2.5753	0.3253
11	138.25	2.939	3.6833	0.7443
12	101.4	1.5531	1.8384	0.2853
13	87.9	2.8297	2.8297	0
14	37.6	1.03	1.1569	0.1269
15	53.8	1.572	1.6267	0.0548
16	67.3	3.0966	3.1161	0.0195
17	13.4	2.0128	2.0413	0.0286
18	104.7	1.3458	1.3458	0
19	58.5	1.526	2.1838	0.6578
20	0	0.1522	0.1522	0
21	27.5	1.055	2.076	1.021

Length in cM: Length of the chromosomal segment covered by the typed markers. Values in **bold** indicate the most promising imprinting test result, for which an empiric p value was calculated using the two proposed MOBIT simulation/permutation procedures.

4 Discussion

Linkage-based testing for genomic imprinting is a challenging task. This holds true for both parametric and nonparametric linkage analysis methods. In this paper, we proposed the likelihood ratio test statistic MOBIT as a new test for imprinting, which is based on the parametric MOD score approach (Clerget-Darpoux, Bonaïti-Pellié & Hochez, 1986; Risch, 1984). The MOBIT is not restricted to the analysis of certain pedigree types, offers quantification of imprinting, does not assume independent parental meioses, and can readily be calculated using the GHM software package (Brugger & Strauch, 2014; Dietter et al., 2007; Mattheisen et al., 2008; Strauch, 2003), which also allows usage of sex-specific maps in the analysis. Although the MOBIT can be considered as a canonical approach to test for imprinting in the presence of linkage, the null distribution of the MOBIT depends on the truly underlying but generally unknown mode of inheritance, i.e. disease allele frequency and penetrance function, which corresponds to a certain point in terms of allele-sharing within the allele-sharing parameter space of a given type of pedigree (see Figure 1 for the example of ASPs). We have shown that the MOBIT asymptotically follows a χ^2 distribution with 1 df irrespective of the pedigree type (see Appendix A.3).

In the special case of no linkage, the MOBIT follows a mixture of distributions that includes non-χ^2 components (see also Self and Liang (1987), case 8, pp. 608–609). As shown in this paper, this leads to quantiles that can be larger than those obtained assuming χ^2 with 1 df (see Table 3 for 3-G pedigrees). Generally, imprinting tests based on linkage test statistics are only advised in the case that linkage can be assumed to be present. But even in the presence of linkage, the quality of the asymptotic approximation of the MOBIT distribution strongly depends on the truly underlying mode of inheritance. If the true point in terms of allele-sharing lies near the boundary of the parameter space, MOBIT quantiles can be smaller or larger than those obtained assuming χ^2 with 1 df. The degree of deviation from the asymptotic distribution directly depends on the structure of the parameter space of a given pedigree. Presumably, allele-sharing parameter spaces become more complicated with increasing pedigree complexity, which in turn means that more boundary conditions are to be expected with more complex pedigree structures. This assumption was underpinned by the present study, such that MOBIT quantiles for the additive trait model and the 1:1 map ratio corresponded to those expected assuming a χ^2 distribution with 1 df when ASPs were used in the analysis, whereas quantiles were clearly inflated when using 3-G pedigrees (Table 4 and Table 5, respectively). To circumvent the uncertainty about the quality of the asymptotic approximation of the MOBIT distribution, we proposed the *ab initio* simulation (method *bfnm*) of genotype data based on the best-fitting nonimprinting model obtained from the real dataset MOD score analysis. Alternatively, we developed a permutation procedure (method *perm*) similar to those proposed in Whittaker et al. (2003) and Dong et al. (2005), which generates replicates under the null hypothesis of no imprinting effect, conditional on the linkage information of the real dataset. During the permutation, the parental origin of transmitted alleles is randomized, while genotypes and hence linkage information are preserved (see Appendix C for more details). We have investigated the distributions of MOBIT scores using both methods *bfnm* and *perm*. However, both methods do not compensate for confounding (see Appendix Table 8). In addition, empiric quantiles and p values differed between the two methods due to differences in the underlying null hypotheses and sample spaces. Using a real data example on house dust mite allergy, both methods to obtain empiric values were compared, which, however, led to different results. This finding indicates that differences in null hypotheses of the two tests may lead to different results and conclusions, especially for the analysis of complex pedigree data.

Another aspect in linkage-based imprinting testing is the confounding between genomic imprinting and sex-specific recombination fractions. This affects both parametric and nonparametric linkage methods. With respect to the MOBIT, type I error rates were increased under the null hypothesis of linkage but no imprinting ($H_{0,b}$) due to confounding of imprinting and sex-specific recombination fractions in a sex-averaged twopoint analysis for both ASPs and 3-G pedigrees (top of Table 4 and Table 5, respectively, map ratios >1). Confounding was more severe for ASPs, the recessive trait model, the larger sample size, and the larger map ratio. In general, confounding is expected to be more severe for trait models with higher allele-sharing, such as the recessive model used for our simulations, because power to detect linkage and hence the potential of confounding is stronger in such a case. The same argument holds true for the larger sample size. In the case of ASPs, the $H_{0,b}$ distribution of the MOBIT for the 1:1 map ratio under the additive and recessive trait models did not show confounding and followed the assumed χ^2 distribution with 1 df. In the case of 3-G pedigrees, the corresponding MOBIT quantiles were inflated for the additive trait model, probably because the true point in terms of allele-sharing lies on the boundary of the parameter space of 3-G pedigrees. In contrast, the MOBIT followed the assumed χ^2 distribution with 1 df for the recessive trait model, which was expected for non-boundary conditions, irrespective of the pedigree type (see Appendix A.3). Using a multipoint analysis avoided confounding as long as the marker spacing was sufficiently dense. We were able to show that a marker spacing of 1 cM between two consecutive markers (marker-trait locus distance 0.5 cM) was sufficient to avoid confounding across both investigated pedigree types and all map ratios (top of Table 4 and Table 5). However, a marker spacing of 10 cM between two consecutive markers (marker-trait locus distance 5 cM) led to confounding, with higher inflated type I error rates for the recessive trait model and for ASPs (top of Table 4 and Table 5).

With regard to the results of the sex-specific analyses (bottom of Table 4 and Table 5), type I error rates of the MOBIT were not inflated due to confounding for both pedigree types and trait models. However, type I error rates were deflated for the MOBIT for the twopoint scenarios and—to a lesser extent—for the 4 markers, 10 cM scenario due to the problem of maximization curves as depicted in Figure 2. Interestingly, the effect of maximization curves was less pronounced for 3-G pedigrees, which indicates differences in the parameter spaces between nuclear families and extended pedigrees. In addition, the effect of maximization curves seemed to be slightly more severe for the recessive trait model, although this was only seen for ASPs (bottom of Table 4).

Apart from the type I error rate, we also assessed the corresponding power of the MOBIT to detect imprinting. As expected, power to detect imprinting was higher for the larger sample size in all scenarios. In the case of a map ratio of 1:1, power was generally higher for ASPs (Figure 4 and Figure 5). With regard to the power calculations when using a sex-averaged map for the analysis, the MOBIT had reasonable power to detect im-

printing across all investigated marker scenarios, map ratios, and pedigree types (Figure 4 and Figure 5). The power of the MOBIT only slightly depended on the truly underlying map ratio for the 1 marker, 5 cM scenario and—to a lesser extent—for the 40 SNPs, 0.32 cM scenario for both pedigree types. Due to the maximization curve problem, the MOBIT had almost no power in the sex-specific twopoint analyses with map ratios >1 using ASPs for the maternal imprinting model (second row of Figure 4). With 3-G pedigrees, however, the MOBIT had markedly better power for all map ratios than with ASPs (last row of Figure 4). With regard to the multipoint results using 40 SNPs, the MOBIT showed good power to detect imprinting for both pedigree types. Using a paternal imprinting model, in which the nonimprinted sex now has the longer genetic map, the power to detect imprinting was reasonably high for sex-averaged and sex-specific analyses as well as for both pedigree types (Figure 5).

Although the power to detect imprinting is bounded from above by the power to detect linkage, this does not imply that the power to detect imprinting must be similar for different pedigrees showing similar power to detect linkage. Put another way, imprinting information is not equivalent to linkage information. Along these lines, ASPs had generally greater power to detect imprinting than the 3-G pedigrees used in our simulations, although the datasets for both pedigree types contained equal amounts of linkage information due to adjusted sample sizes. This is because the 3-G pedigree was constructed such that, for maternal imprinting, a considerable proportion of phenocopies and/or carriers of maternally inherited mutant alleles were simulated as affected individuals 8 and 9, because of the mother individual 5 propagating the mutant allele down the pedigree (see Figure 3B). The mutant allele might also have entered the pedigree through founder individual 3 (spouse of mother individual 5), although this being less likely in regard of the rather small disease allele frequency of $p = 0.1$.

Using MOD scores, it should in principle be possible to obtain unbiased estimates of the trait-model parameters f_0, $f_{1,pat}$, $f_{1,mat}$, f_2, and p (Elston, 1989). This is due to the fact that the likelihood ratio in the MOD score corresponds to the conditional probability of observing the marker data given the trait phenotypes (Clerget-Darpoux, Bonaïti-Pellié & Hochez, 1986). However, the identifiability of the trait-model parameters actually depends on the number of free parameters in terms of allele-sharing classes of the investigated pedigree type(s). With ASPs, the allele-sharing classes are z_0, z_1^{pat}, z_1^{mat}, and z_2 when taking imprinting into account. Hence, as there are only $4 - 1 = 3$ free parameters that can be estimated from ASP data, there will be many sets of f_0, $f_{1\,pat}$, $f_{1\,mat}$, f_2, p, and θ that correspond to the estimated z_0, z_1^{pat}, z_1^{mat}, and z_2. With larger pedigrees, and hence more allele-sharing classes, the degree to which the trait-model parameters can be determined should be higher. Further information reffering to the ability of a MOD score analysis to correctly determine the truly underlying trait-model parameters for different pedigree types can be found in Brugger, Rospleszcz, and Strauch (2016). Here, we were interested in the ability of a MOBIT analysis to correctly estimate the degree of imprinting, as defined by the imprinting index I (Strauch, 2005). As a result, the estimated median imprinting indices of the two- and multipoint analyses assuming a sex-averaged map were close to their expected values for ASPs and the maternal imprinting model, except the 1 marker, 5 cM scenario, for which indices were overestimated when <0.6 and underestimated when ≥ 0.6 (Figure 6). In the case of 3-G pedigrees, imprinting indices for the maternal imprinting model were mostly close to their expected values, although always underestimated (Figure 7). In the case of the paternal imprinting model, estimates of the imprinting index were often underestimated for both pedigree types due to the longer genetic map of the nonimprinted sex (Figure 5). The results of the sex-specific analyses for ASPs and the maternal imprinting model in Figure 8 indicated that the estimates were biased towards lower values due to the maximization curve problem, which is in line with the reduced power values shown in Figure 4. The estimates for the 3-G pedigrees, however, were only slightly different from those of the corresponding sex-averaged analysis (Figure 9). Specifically, estimated median imprinting indices were again always underestimated, but to a slightly more severe degree than it was for the sex-averaged analysis. In contrast, in the case of the paternal imprinting model, imprinting indices could be estimated more accurately using the correct sex-specific map in the analysis for both pedigree types (Figure 5). Generally, imprinting indices could be estimated more accurately using ASP pedigrees, because they harbour more imprinting information compared to the 3-G pedigrees studied here, although power to detect linkage was similar for both pedigree types (see also explanation above).

Referring to the question whether to opt for a two- or multipoint approach in a MOBIT imprinting analysis, multipoint analysis should generally be preferred, because it showed good power irrespective of the truly underlying map ratio and whether or not the correct sex-specific map was used in the analysis (see Figure 4). Moreover, quantification of imprinting in terms of the imprinting index I is more reliable in the multipoint setting, especially when marker spacing is dense.

It is of note, however, that differences in heterozygote penetrances might also be caused by another parent-of-origin effect, namely maternal effects. Maternal effects refer to the phenotype of an individual being influenced by the genotype of the mother (Han, Hu & Lin, 2013), whereby it can induce the same phenotypic pattern in the offspring as genomic imprinting (Hager, Cheverud & Wolf, 2008). Hence, maternal effects and genomic

imprinting are confounded and cannot be distinguished using the MOBIT or any other existing nonparametric linkage method by looking at different heterozygote penetrances or parental allele-sharing values, respectively. A way to separate maternal effects from genomic imprinting has been proposed for parent-offspring data and was investigated in the context of quantitative trait loci by Hager, Cheverud, and Wolf (2008). An extension of the MOD score approach that allows to distinguish maternal effects from imprinting would be an interesting future research item, because this would further refine the possibility of the MOD score approach to characterize the disease gene variant, based on trait-model parameter estimates as shown by Brugger, Rospleszcz, and Strauch (2016).

In this work, we have proposed the new imprinting test statistic MOBIT and evaluated its properties using extensive simulations. With regard to the effect of confounding, the MOBIT showed inflated type I error rates when the marker spacing was not dense enough (> 1 cM), which can be remedied by using sex-specific maps. However, sex-specific twopoint analyses should be avoided, because the test for imprinting has no power when the underlying genetic map ratio is large, which is due to the so-called maximization curve problem. When a sufficiently dense marker map with a marker spacing of less than 1 cM is used in the analysis, the difference between sex-specific and sex-averaged maps is marginal, if not negligible. In such a case, the MOBIT did not show an effect of confounding, had good power to detect imprinting, and was able to reliably estimate the degree of imprinting, the accuracy of which depended on the pedigree type used in the analysis. Hence, we recommend the usage of sufficiently dense marker frameworks to avoid both confounding of sex-specific recombination fractions and imprinting as well as low power due to the maximization curve problem. In addition, we proposed two alternative simulation/permutation methods to obtain empiric p values for the MOBIT and compared them using a real dataset and various scenarios of the main simulation study. In general, we recommend to apply the method *bfnm*, which relies on the fully parametric *ab initio* simulation of genotypes according to the best-fitting nonimprinting model obtained from the real dataset analysis, because it is the canonical approach to generate replicates under the null hypothesis of linkage, but no imprinting. Replicates for the method *bfnm* can readily be generated using the SLINK software package (Ott, 1989; Weeks et al., 1990; Schäffer et al., 2011). However, there might be situations in which the null hypothesis of method *perm*, i.e. no imprinting effect, conditional on the linkage information of the real dataset, might also prove useful, e.g. when the research interest lies in testing a null hypothesis that is confined to the exact realisation of the real dataset. We implemented the *perm* method in a new version of the GHM software package. Taken together, the MOBIT is a recommendable tool to discover new imprinted loci and offers good flexibility with regard to marker scenarios and pedigree types.

The imprinting test statistic MOBIT and a corresponding p value according to the *perm* method can be calculated using the GENEHUNTER-MODSCORE program, which has recently been optimized to provide a fast linkage analysis with evaluation of many sets of trait-model parameters by algebraic computations (Brugger & Strauch, 2014). The program can be freely downloaded from our website http://www.helmholtz-muenchen.de/ige/service/software-download/index.html.

Acknowledgement

This work was supported by grants Str643/4-1 and Str643/6-1 of the Deutsche Forschungsgemeinschaft (German Research Foundation). Further, this research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this work by providing computing time on the GCS Supercomputer SuperMUC at Leibniz Supercomputing Centre (www.lrz.de). In addition, the authors are grateful for the computing resources on the High Performance Computing – High Availability – Cluster (HPC-HA-Cluster) provided by the Helmholtz Zentrum München. Finally, we would like to thank the reviewers for their thoughtful comments that helped us to substantially improve the manuscript.

Funding

Deutsche Forschungsgemeinschaft, Funder Id: http://dx.doi.org/10.13039/501100001659, Grant Number: Str643/4-1 and Str643/6-1.

A Appendix

A.1 Proof of the equivalence of the nonimprinting and imprinting likelihood in the test for imprinting using affected sib-pairs (ASPs) in the case of no linkage

An ASP with arbitrary genotype information (missing or non-missing) for parents and offspring is given. Without loss of generality, allele sharing at a single marker is considered. The following equality is to be shown:

$$\prod_{i=1}^{n}\left(w_{i0}\cdot\frac{1}{4}+w_{i1}\cdot\frac{1}{2}+w_{i2}\cdot\frac{1}{4}\right)=\prod_{i=1}^{n}\left(w_{i0}\cdot\frac{1}{4}+w_{i1^{pat}}\cdot\frac{1}{4}+w_{i1^{mat}}\cdot\frac{1}{4}+w_{i2}\cdot\frac{1}{4}\right) \tag{3}$$

with

$$w_{ij}=P(g_{\text{off},\,i}|ibd_i=j)=\frac{P(ibd_i=j|g_{\text{off},\,i})\cdot P(g_{\text{off},\,i})}{P(ibd_i=j)} \tag{4}$$

- $P(g_{off,\,i})$: probability of genotypes of both offspring for the i-th ASP
- pat/mat: determines parental, i.e. paternal or maternal, origin of the shared allele in the imprinting likelihood formulation
- $P(ibd_i=j)$: probability of sharing $j=0,\,1,\,2$ (or $0,\,1^{pat},\,1^{mat}$,2 with imprinting) alleles identical-by-descent (IBD) for the i-th ASP under the null hypothesis of no linkage, which is $\left\{\frac{1}{4},\,\frac{1}{2},\,\frac{1}{4}\right\}$ for the nonimprinting and $\left\{\frac{1}{4},\,\frac{1}{4},\,\frac{1}{4},\,\frac{1}{4}\right\}$ for the imprinting likelihood, respectively
- $P(ibd_i=j\,|\,g_{off,\,i})$: probability of the i-th ASP sharing $j=0,\,1,\,2$ (or $0,\,1^{pat},\,1^{mat}$,2 with imprinting) alleles IBD, conditional on offspring genotypes, which is $\{z_{i0},\,z_{i1},\,z_{i2}\}$ for the nonimprinting and $\left\{z_{i0},\,z_{i1}^{pat},\,z_{i1}^{mat},\,z_{i2}\right\}$ for the imprinting likelihood, respectively

In order to show that equation (3) holds, it is sufficient to show that the corresponding equality holds for each ASP:

$$w_{i0}\cdot\frac{1}{4}+w_{i1}\cdot\frac{1}{2}+w_{i2}\cdot\frac{1}{4}=w_{i0}\cdot\frac{1}{4}+w_{i1^{pat}}\cdot\frac{1}{4}+w_{i1^{mat}}\cdot\frac{1}{4}+w_{i2}\cdot\frac{1}{4} \tag{5}$$

Given the assumption that w_{i0} and w_{i2} do not change when the case of $ibd=1$ shared allele is distinguished according to its parental origin, this reduces to:

$$\begin{aligned} w_{i1}\cdot\frac{1}{2}&=w_{i1^{pat}}\cdot\frac{1}{4}+w_{i1^{mat}}\cdot\frac{1}{4}\\ w_{i1}&=\frac{w_{i1^{pat}}+w_{i1^{mat}}}{2}\end{aligned} \tag{6}$$

Using equation (4), we can write for the right side of equation (6):

$$\begin{aligned} w_{i1^{pat}}+w_{i1^{mat}}&=\frac{P(ibd_1=1^{pat}|g_{\text{off},\,1})\cdot P(g_{\text{off},\,1})}{\frac{1}{4}}\\ &\quad+\frac{P(ibd_1=1^{mat}|g_{\text{off},\,1})\cdot P(g_{\text{off},\,1})}{\frac{1}{4}}\\ &=4\cdot P(g_{\text{off},\,1})\cdot\left\{P(ibd_1=1^{pat}|g_{\text{off},\,1})+P(ibd_1=1^{mat}|g_{\text{off},\,1})\right\}\\ &=4\cdot P(g_{\text{off},\,1})\cdot P(ibd_1=1|g_{\text{off},\,1})\end{aligned} \tag{7}$$

The last step holds because $z_{11}=z_{11}^{pat}+z_{11}^{mat}$. For w_{i1}, we get:

$$\begin{aligned} w_{i1}&=\frac{P(ibd_1=1|g_{\text{off},\,1})\cdot P(g_{\text{off},\,1})}{\frac{1}{2}}\\ &=2\cdot P(g_{\text{off},\,1})\cdot P(ibd_1=1|g_{\text{off},\,1})\end{aligned} \tag{8}$$

Replacing last line of equation (7) and last line of equation (8) into the last line of equation (6), we get:

$$2 \cdot \left\{2 \cdot P(g_{\text{off},\,1}) \cdot P(ibd_1 = 1|g_{\text{off},\,1})\right\} = 4 \cdot P(g_{\text{off},\,1}) \cdot P(ibd_1 = 1|g_{\text{off},\,1}) \tag{9}$$

Since this holds in general for each ASP, it follows that equation (3) holds for the entire dataset with any number of ASPs.

A.2 Proof of the equivalence of the nonimprinting and imprinting likelihood in the test for imprinting using arbitrary pedigrees in the case of no linkage

An arbitrary pedigree type with arbitrary genotype information (missing or non-missing) for founders and nonfounders is given. Without loss of generality, allele-sharing at a single marker is considered. The numbers of shared alleles IBD for affected relative pairs are generalized to allele-sharing classes for arbitrary pedigree structures (see also Brugger and Strauch (2014)). The following equality is to be shown:

$$\prod_{i=1}^{n} \sum_{j=0}^{j \in \mathcal{V}_{nonimp}} \left(w_{ij} \cdot P(ibd_i = j)\right) = \prod_{i=1}^{n} \sum_{j=0}^{j \in \mathcal{V}_{imp}} \left(w_{ij} \cdot P(ibd_i = j)\right) \tag{10}$$

$$w_{ij} = P(g_{\text{off},\,i}|ibd_i = j) = \frac{P(ibd_i = j|g_{\text{off},\,i}) \cdot P(g_{\text{off},\,i})}{P(ibd_i = j)} \tag{11}$$

with $\mathcal{V}_{nonimp}$ denoting the set of allele-sharing classes without imprinting and $\mathcal{V}_{imp}$ denoting the set of allele-sharing classes with imprinting.

- $P(g_{off,\,i})$: probability of genotypes of all offsprings for the i-th pedigree
- $P(ibd_i = j)$: probability of allele-sharing class j for the i-th pedigree under the null hypothesis of no linkage, which is a constant value
- $P(ibd_i = j \mid g_{off,\,i})$: probability of allele-sharing class j for the i-th pedigree, conditional on offspring genotypes

In order to show that equation (10) holds, it is sufficient to show that the corresponding equality holds for each pedigree:

$$\sum_{j=0}^{j \in \mathcal{V}_{nonimp}} \left(w_{ij} \cdot P(ibd_i = j)\right) = \sum_{j=0}^{j \in \mathcal{V}_{imp}} \left(w_{ij} \cdot P(ibd_i = j)\right) \tag{12}$$

For those allele-sharing classes that are identical with and without imprinting, i.e. those contained in $\mathcal{V}_{nonimp} \cap \mathcal{V}_{imp}$, the corresponding w_{ij} are identical. Hence, equation (12) reduces to:

$$\sum_{w \in \mathcal{V}_{nonimp}\ (\mathcal{V}_{nonimp} \cap \mathcal{V}_{imp})} \left(w_{ij} \cdot P(ibd = j)\right) = \sum_{w \in \mathcal{V}_{imp}\ (\mathcal{V}_{nonimp} \cap \mathcal{V}_{imp})} \left(w_{ij} \cdot P(ibd = j)\right) \tag{13}$$

We now plug in equation (11) into equation (13) to get the following:

$$\begin{aligned} &\sum_{w \in \mathcal{V}_{nonimp}\ (\mathcal{V}_{nonimp} \cap \mathcal{V}_{imp})} \left(\frac{P(ibd = j|g_{\text{off}}) \cdot P(g_{\text{off}})}{P(ibd = j)} \cdot P(ibd = j)\right) \\ &= \sum_{w \in \mathcal{V}_{imp}\ (\mathcal{V}_{nonimp} \cap \mathcal{V}_{imp})} \left(\frac{P(ibd = j|g_{\text{off}}) \cdot P(g_{\text{off}})}{P(ibd = j)} \cdot P(ibd = j)\right) \end{aligned} \tag{14}$$

$$\begin{aligned} &\sum_{w \in \mathcal{V}_{nonimp}\ (\mathcal{V}_{nonimp} \cap \mathcal{V}_{imp})} \left(P(ibd = j|g_{\text{off}}) \cdot P(g_{\text{off}})\right) \\ &= \sum_{w \in \mathcal{V}_{imp}\ (\mathcal{V}_{nonimp} \cap \mathcal{V}_{imp})} \left(P(ibd = j|g_{\text{off}}) \cdot P(g_{\text{off}})\right) \end{aligned} \tag{15}$$

$$P(g_{\text{off}}) \cdot \sum_{w \in \mathscr{V}_{nonimp}} \sum_{(\mathscr{V}_{nonimp} \cap \mathscr{V}_{imp})} P(ibd = j|g_{\text{off}}) = P(g_{\text{off}}) \cdot \sum_{w \in \mathscr{V}_{imp}} \sum_{(\mathscr{V}_{nonimp} \cap \mathscr{V}_{imp}),} P(ibd = j|g_{\text{off}}) \tag{16}$$

$$\sum_{w \in \mathscr{V}_{nonimp}} \sum_{(\mathscr{V}_{nonimp} \cap \mathscr{V}_{imp})} P(ibd = j|g_{\text{off}}) = \sum_{w \in \mathscr{V}_{imp}} \sum_{(\mathscr{V}_{nonimp} \cap \mathscr{V}_{imp})} P(ibd = j|g_{\text{off}}) \tag{17}$$

Given that the sum of IBD probabilities of non-overlapping allele-sharing classes without imprinting must equal the corresponding sum with imprinting, the likelihoods are identical. This also extends to datasets containing mixtures of different pedigree types.

A.3 Proof that the MOBIT follows a χ^2 distribution with 1 degree of freedom under the null hypothesis of linkage but no imprinting for arbitrary pedigrees

Given the proof in Appendix A.2, the MOBIT for arbitrary pedigree types, corresponding to equation (1) in the Methods section for the special case of ASPs, can also be written as a likelihood ratio of the nonparametric likelihoods taking imprinting into account (numerator) vs. not taking imprinting into account (denominator). The nonparametric allele-sharing probabilities $\mathbf{z}$ can be further expressed as functions of the penetrances, the disease allele frequency, and the recombination fraction (see also Brugger and Strauch (2014)):

$$\begin{aligned} MOBIT &= \log_{10} \frac{L\left(\hat{\mathbf{z}}_{\mathbf{imp}}\right)}{L\left(\hat{\mathbf{z}}_{\mathbf{nonimp}}\right)} \\ &= \log_{10} \frac{L\left(\mathbf{z}_{\mathbf{imp}}(\hat{f}_0, \hat{f}_{1\,pat}, \hat{f}_{1\,mat}, \hat{f}_2, \hat{p}, \hat{\theta})\right)}{L\left(\mathbf{z}_{\mathbf{nonimp}}(\hat{f}_0, \hat{f}_1, \hat{f}_2, \hat{p}, \hat{\theta})\right)} \\ &= \log_{10} \frac{L\left(\hat{f}_0, \hat{f}_{1\,pat}, \hat{f}_{1\,mat}, \hat{f}_2, \hat{p}, \hat{\theta}\right)}{L\left(\hat{f}_0, \hat{f}_1, \hat{f}_2, \hat{p}, \hat{\theta}\right)} \end{aligned} \tag{18}$$

, where $\hat{\mathbf{z}}_{\mathbf{imp}}$, $\hat{\mathbf{z}}_{\mathbf{nonimp}}$, $\hat{f}_0$, $\hat{f}_1$, $\hat{f}_{1\,pat}$, $\hat{f}_{1\,mat}$, $\hat{f}_2$, $\hat{p}$, $\hat{\theta}$ denote the maximum likelihood estimators (MLEs) of the respective parameters. Likelihood ratio theory tells us that the ratio in equation (18) follows a χ^2 distribution with the number of degrees of freedom equal to the difference of independently maximized parameters in the numerator vs. the denominator (Wilks, 1938), which equals 1 in our case. Further, the hypothesis in the denominator should be a nested composite hypothesis of the one in the numerator (Wilks, 1938), which is true, because fixing $f_{1,\,pat} = f_{1,\,mat}$ in the numerator corresponds to the likelihood in the denominator. It is of note, however, that the asymptotic distribution no longer holds in the case of boundary conditions with respect to the alternative hypothesis in the numerator (Self & Liang, 1987). In such a case, the empiric MOBIT distribution quantiles can be lower or higher than those expected for a χ^2_1 distribution.

B Appendix

B.1 Proof that the marker-trait locus distance is identifiable in a MOD score analysis using ASPs if: i) sex-specific recombination frequencies are present and used in the analysis; and ii) imprinting is truly absent

Let $\left(\tilde{f}_2, \tilde{f}_{1,\,pat}, \tilde{f}_{1,\,mat}, \tilde{f}_0, \tilde{p}\right)$ and $\left(f_2^*, f_{1\,pat}^*, f_{1\,mat}^*, f_0^*, p^*\right)$ be two genetic disease models without imprinting, i.e. $\tilde{f}_{1,\,pat} = \tilde{f}_{1,\,mat}$ and $f_{1\,pat}^* = f_{1\,mat}^*$, and $\tilde{u} = \left(\tilde{u}_2, \tilde{u}_1^{pat}, \tilde{u}_1^{mat}, \tilde{u}_0\right) \neq (1/4,\ 1/4,\ 1/4,\ 1/4)$ and $u^* =$

$\left(u_2^*,\ u_1^{*pat},\ u_1^{*mat},\ u_0^*\right) \neq (1/4,\ 1/4,\ 1/4,\ 1/4)$ the corresponding IBD distributions at the disease locus with $\tilde{u}_1^{pat} = \tilde{u}_1^{mat}$ and $u_1^{*pat} = u_1^{*mat}$. Further, $\tilde{\theta}_m \neq \tilde{\theta}_f$ and $\theta_m^* \neq \theta_f^*$ are sex-specific recombination frequencies (m: male; f: female) and $\tilde{z}$ and z^* are the IBD distributions at the marker locus that are induced by the sex-specific recombination frequencies. Proposition: From $\tilde{z} = z^*$, it follows:

$$C_f^* = \frac{\tilde{C}_f}{\tilde{C}_m} C_m^* \tag{19}$$

with $\tilde{C}_f := \tilde{\theta}_f^2 + (1-\tilde{\theta}_f)^2 - 0.5$, $\tilde{C}_m := \tilde{\theta}_m^2 + (1-\tilde{\theta}_m)^2 - 0.5$ and $C_f^* := \theta_f^{*2} + (1-\theta_f^*)^2 - 0.5$, $C_m^* := \theta_m^{*2} + (1-\theta_m^*)^2 - 0.5$

Proof:

With

$$T(\psi_m,\ \psi_f) = \begin{pmatrix} \psi_m\psi_f & \psi_m(1-\psi_f) & (1-\psi_m)\psi_f & (1-\psi_m)(1-\psi_f) \\ \psi_m(1-\psi_f) & \psi_m\psi_f & (1-\psi_m)(1-\psi_f) & (1-\psi_m)\psi_f \\ (1-\psi_m)\psi_f & (1-\psi_m)(1-\psi_f) & \psi_m\psi_f & \psi_m(1-\psi_f) \\ (1-\psi_m)(1-\psi_f) & (1-\psi_m)\psi_f & \psi_m(1-\psi_f) & \psi_m\psi_f \end{pmatrix}$$

we have $\tilde{z} = T(\tilde{\psi}_m,\ \tilde{\psi}_f) \cdot \tilde{u}$ and hence

$$\begin{aligned} \tilde{z}_1^{pat} - \tilde{z}_1^{mat} &= (\tilde{\psi}_m - \tilde{\psi}_f) \cdot (\tilde{u}_2 - \tilde{u}_0), \\ \tilde{z}_2 - \tilde{z}_0 &= (\tilde{\psi}_m + \tilde{\psi}_f - 1) \cdot (\tilde{u}_2 - \tilde{u}_0). \end{aligned}$$

Let now be $S(\psi_m,\ \psi_f) =$

$$\begin{pmatrix} \psi_m\psi_f & -\psi_m(1-\psi_f) & -(1-\psi_m)\psi_f & (1-\psi_m)(1-\psi_f) \\ -\psi_m(1-\psi_f) & \psi_m\psi_f & (1-\psi_m)(1-\psi_f) & -(1-\psi_m)\psi_f \\ -(1-\psi_m)\psi_f & (1-\psi_m)(1-\psi_f) & \psi_m\psi_f & -\psi_m(1-\psi_f) \\ (1-\psi_m)(1-\psi_f) & -(1-\psi_m)\psi_f & -\psi_m(1-\psi_f) & \psi_m\psi_f \end{pmatrix}$$

It can be shown that $T^{-1}(\tilde{\psi}_m,\ \tilde{\psi}_f) = \frac{1}{(\psi_m^2-(1-\psi_m)^2)\cdot(\psi_f^2-(1-\psi_f)^2)} \cdot S(\psi_m,\ \psi_f)$. Due to $\tilde{z} = z^*$ we have $u^* = T^{-1}(\psi_m^*,\ \psi_f^*) \cdot \tilde{z}$ and hence also

$$\begin{aligned} 0 = u_1^{*pat} - u_1^{*mat} &= (\psi_f^* - \psi_m^*) \cdot (\tilde{z}_2 - \tilde{z}_0) + (\psi_f^* + \psi_m^* - 1) \cdot (\tilde{z}_1^{pat} - \tilde{z}_1^{mat}) \\ &= \left[(\psi_f^* - \psi_m^*) \cdot (\tilde{\psi}_m + \tilde{\psi}_f - 1) - (\psi_f^* + \psi_m^* - 1) \cdot (\tilde{\psi}_f - \tilde{\psi}_m) \right] \cdot (\tilde{u}_2 - \tilde{u}_0) \\ &= \left[\psi_f^* \cdot (2\,\tilde{\psi}_m - 1) - \psi_m^* \cdot (2\,\tilde{\psi}_f - 1) + (\tilde{\psi}_f - \tilde{\psi}_m) \right] \cdot (\tilde{u}_2 - \tilde{u}_0) \\ &= 2(C_f^* \cdot \tilde{C}_m - C_m^* \cdot \tilde{C}_f) \cdot (\tilde{u}_2 - \tilde{u}_0) \end{aligned}$$

Due to $u \neq (1/4,\ 1/4,\ 1/4,\ 1/4)$ we have $\tilde{u}_2 - \tilde{u}_0 > 0$, which leads to proposition (19). It remains to be shown that for two pairs $(\tilde{\theta}_m,\ \tilde{\theta}_f)$ and $(\theta_m^*,\ \theta_f^*)$, for which equation (19) holds and for which the ratios of genetic map distances (using the Haldane mapping function) are identical, $\tilde{\theta}_m = \theta_m^*$ and $\tilde{\theta}_f = \theta_f^*$ is always true. To this end, it should be noted that the range of values for $C(\theta) = \theta^2 + (1-\theta)^2 - 0.5$ corresponds to the interval $(0,\ 1/2)$ if $\theta \in (0,\ 1/2)$ and that $C((1-\sqrt{2U})/2) = U$. Hence, for the genetic distance x_f^* corresponding to θ_f^* we have

$$\begin{aligned} x_f^* &= -0.5 \ln\left(1 - 2 \cdot \left(1 - \sqrt{2C_f^*}\right)/2\right) = -0.25 \log(2C_f^*) \text{and} \\ \frac{x_f^*}{x_m^*} &= \frac{\log 2C_f^*}{\log 2C_m^*}. \end{aligned}$$

With $\tilde{a} := \tilde{C}_f \,/\, \tilde{C}_m$ we hence get

$$\frac{x_f^*}{x_m^*} = \frac{\log 2\ \tilde{a}\ C_m^*}{\log 2 C_m^*}$$

if $(\theta_m^*,\ \theta_f^*)$ fulfill equation (19).

In the same line,

$$\frac{\tilde{x}_f}{\tilde{x}_m} = \frac{\log 2a^*\ \tilde{C}_m}{\log 2\ \tilde{C}_m}$$

holds with $a^* := C_f^*/C_m^*$ and $\tilde{a} = a^* := a$ due to equation (19).

If then

$$\frac{x_f^*}{x_m^*} = \frac{\tilde{x}_f}{\tilde{x}_m},$$

it follows from the strict montony of the function $g(v) = \frac{\log(2av)}{\log(2v)}$ for $v \in (0,\ 1/2)$ that $\tilde{C}_m = C_m^*$ and also, due to equation (19), $\tilde{C}_f = C_f^*$. Finally, due to the strict monotony of the function $C(\theta)$, it follows $\tilde{\theta}_m = \theta_m^*$ and $\tilde{\theta}_f = \theta_f^*$.

Furthermore, the identifiability of the disease locus position for ASPs in a sex-specific MOD score analysis could experimentally be confirmed using the scenario of the main simulation study with 5 cM marker-trait locus distance (Appendix Table 7). The results further indicated that the position is also identifiable using 3-G pedigrees, even in the case of no sex-specific differences in the recombination fraction.

Table 7: Estimated median marker-trait locus distances under $H_{0,b}$: Linkage, no imprinting when using the sex-specific map in the nonimprinting analysis as employed for the simulation.

True marker-trait locus distance: 5 cM	**Estimated median marker-trait locus distance (MAD)**			
Map ratio		**600 ASPs**		**65 3-G pedigrees**
	additive	**recessive**	**additive**	**recessive**
1:1	21 (14.83)	9 (5.93)	3 (4.45)	5 (7.41)
7:3	5 (7.41)	5 (4.45)	4 (5.93)	5 (7.41)
9:1	5 (7.41)	5 (2.97)	4 (5.93)	5 (4.45)

MAD: median absolute deviation, adjusted by a constant (1.4826) for asymptotically normal consistency. ASP: affected sib-pair; 3-G pedigree: three-generation pedigree.

C Appendix

C.1 Description and results of the newly developed MOBIT permutation procedure (method *perm*)

In brief, replicates under the null hypothesis of an expected imprinting effect of 0, conditional on the linkage information of the real dataset, are generated by leaving the marker genotypes of the real dataset untouched, while the parental origin of the alleles in the offspring is randomly permuted for each permutation unit in a given pedigree. Here, a permutation unit is defined as a nuclear family or, in the case of half-sibs, as a set of nuclear families, whose parents are connected by matings with joint offspring. After identification of permutation units, randomization of imprinting information is achieved by manipulation of the parametric disease-locus likelihood, where each inheritance vector in the Lander-Green algorithm (Lander & Green, 1987), on which GHM is based, is assigned a score as a function of the disease allele frequency and the penetrance values (see also Brugger and Strauch (2014) for more information on the Lander-Green algorithm and the disease-locus likelihood as implemented in GHM). Each inheritance vector score is then weighted by the corresponding probability of the marker genotype data given the particular inheritance vector. It can be shown that certain

inheritance vectors can be collapsed into classes on the basis of the equality of their disease-locus likelihood contribution (Brugger & Strauch, 2014). In the case of imprinting, where two heterozygote penetrances are modeled, there is a larger number of inheritance vector classes than without imprinting. This is because certain inheritance vector classes without imprinting split up in two or more classes by replacing each factor of the heterozygote penetrance (f_1) by the corresponding parental-origin-specific value ($f_{1\,pat}$ or $f_{1\,mat}$). It is these inheritance vector classes that get randomly permuted for a given permutation unit of a pedigree in each new replicate. It is of note, however, that confounding between imprinting and sex-specific recombination fractions cannot be avoided using this permutation procedure if the truly underlying sex-specific maps are not used in the analysis. If sex-specific maps are used in the analysis, they are correctly handled by the permutation procedure. In order to investigate the properties of the MOBIT permutation distribution, we picked a dataset from the main simulation study and calculated the corresponding permutation distribution as well as the p value for this dataset. An overview of the investigated scenarios can be found in Appendix Table 8. In the case of ASPs, we also graphically depicted the resulting points in terms of allele-sharing for the permuted replicates in the tetrahedral parameter space (Appendix Figure 10 and Appendix Figure 11). This illustrates that the *perm* method described here leads to a sample space that is different from the one obtained for the *bfnm* method, which is similar to those depicted in Figure 1.

Table 8: Overview of the simulated scenarios to investigate the properties of the two proposed MOBIT simulation/permutation procedures.

	Overview of the investigated scenarios for the evaluation of the MOBIT simulation/permutation procedures					
Hypothesis	$H_{0,a}$**: No linkage, no imprinting**		$H_{0,b}$**: Linkage, no imprinting**		H_1**: Linkage, imprinting**	
Trait model	additive/recessive				$I = 0.4$	
Map ratio	1:1	9:1	1:1	9:1	1:1	9:1
Analysis type	sex-averaged	sex-specific	sex-averaged	sex-averaged/sex-specific	sex-averaged	sex-specific
Pedigree type	Sample size	Number of replicates				
ASP	600	10,000				
3-G pedigree	65	10,000				
Scenario: 1 marker, 5 cM (sex-averaged) between markerand disease locus						
Segregation of additive trait simulated with penetrances $\{f_0, f_1, f_2\} = \{0.03,\ 0.13,\ 0.23\}$ and disease allele frequency $p = 0.1$						
Segregation of recessive trait simulated with penetrances $\{f_0, f_1, f_2\} = \{0.05,\ 0.05,\ 0.90\}$ and disease allele frequency $p = 0.2$						
Maternal imprinting simulations for additive trait model with $I = \frac{f_{1\,pat} - f_{1\,mat}}{f_2 - f_0} = 0.4$						

ASP: affected sib-pair; 3-G pedigree: three-generation pedigree.

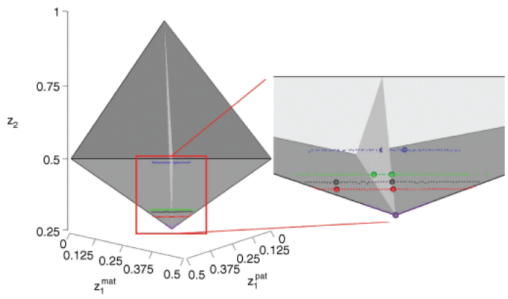


Figure 10: Results of the MOBIT permutation procedure (method *perm*) using a sex-averaged map for the analysis. The illustration is based on the tetrahedral parameter space **T** for affected sib-pairs (ASPs, see also Figure 1). Left subfigure: total view of **T**; right subfigure: detailed view (slightly rotated to the right) on the points in terms of allele-sharing. For more details as to the investigated hypotheses and trait models see Appendix Table 8 and Appendix Table 9. Bullets correspond to points in terms of allele-sharing according to the trait-model parameters and the recombinaion fraction obtained from the MOD (bullets on the possible triangle) and IMOD score analyses of the picked replicate, which was used as the real dataset. In some cases, the position of the MOD and IMOD score bullets cannot be visibly distinguished. Points correspond to allele-sharing IMOD score estimates obtained for each of the 10,000 generated replicates of the MOBIT permutation procedure. The color scheme of bullets and dots is as follows. Purple: no linkage, no imprinting ($H_{0,a}$) with 1:1 map ratio; green: linkage, no imprinting ($H_{0,b}$, additive trait model) with 5 cM marker-trait locus distance and 1:1 map ratio; blue: $H_{0,b}$ (recessive trait model) with 5 cM marker-trait locus distance and 1:1 map ratio; red: $H_{0,b}$ (additive trait model) with 5 cM marker-trait locus distance and 9:1 map ratio; grey: linkage and imprinting (H_1) with 5 cM marker-trait locus distance and 1:1 map ratio.

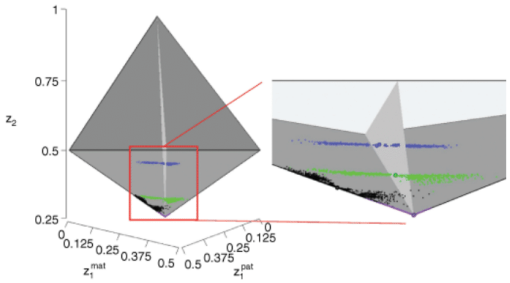


Figure 11: Results of the MOBIT permutation procedure (method *perm*) using the correct sex-specific map with a 9:1 female/male map ratio for the analysis. The illustration is based on the tetrahedral parameter space **T** for affected sib-pairs (ASPs, see also Figure 1). Left subfigure: total view of **T**; right subfigure: detailed view (slightly rotated to the right) on the points in terms of allele-sharing. Bullets correspond to points in terms of allele-sharing according to the trait-model parameters and the recombinaion fraction obtained from the MOD and IMOD score analyses of the picked replicate, which was used as the real dataset. The left of the two blue bullet points corresponds to the best-fitting nonimprinting model obtained from the MOD score analysis. In all other cases, the position of the MOD and IMOD score bullets cannot be visibly distinguished. Points correspond to allele-sharing IMOD score estimates obtained for each of the 10,000 generated replicates of the MOBIT permutation procedure. The color scheme of bullets and dots is as follows. Purple: no linkage, no imprinting ($H_{0,a}$) with 9:1 map ratio; green: linkage, no imprinting ($H_{0,b}$, additive trait model) with 5 cM marker-trait locus distance and 9:1 map ratio; blue: $H_{0,b}$ (recessive trait model) with 5 cM marker-trait locus distance and 9:1 map ratio; grey: linkage and imprinting (H_1) with 5 cM marker-trait locus distance and 9:1 map ratio.

Table 9: Empiric quantiles of the MOBIT distribution for ASPs (top) and 3-G pedigrees (bottom) using the two proposed simulation/permutation procedures.

ASPs																			
	Simulation/analysis using a sex-averaged map									Simulation/analysis using a sex-specific map									
Hypothesis	$H_{0,a}$	$H_{0,b}$					H_1			$H_{0,a}$	$H_{0,b}$					H_1			
Trait model	–	additive				recessive	$I = 0.4$			–	additive			recessive		$I = 0.4$			
Map ratio	1:1	1:1		9:1		1:1	1:1			9:1	9:1			9:1		9:1			
$MOBIT_{real}$	0.0	0.1204		1.1615		0.2856	1.1694			0.0	0.0			0.0063		0.0001			
	Quantiles									Quantiles									
Method	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*
50%	0.0168	0.0	0.0927	0.0862	0.0912	0.1042	0.0999	0.0883	0.0971	0.1049	0.0002	0.0	0.0005	0.0994	0.0023	0.1055	0.0016	0.0948		
90%	0.31	0.0709	0.576	0.5995	0.5683	0.5683	0.5824	0.5971	0.588	0.5721	0.1448	0.0601	0.1313	0.6002	0.0842	0.5973	0.0286	0.5996		
95%	0.4693	0.1173	0.8258	0.7766	0.822	0.8385	0.8398	0.7964	0.822	0.8442	0.2838	0.1216	0.2912	0.8632	0.2173	0.85	0.0558	0.8688		
99%	0.9208	0.2838	1.4281	1.447	1.3878	1.4063	1.4725	1.4185	1.4042	1.4154	0.6709	0.2964	0.7308	1.4806	0.6117	1.4574	0.1058	1.4714		
p value	1	1	0.4503	0.4561	0.019	0.0243	0.2484	0.2548	0.0197	0.0244	1	1	1	1	0.2909	0.8181	0.6904	0.9459		
3-G pedigrees																				
	Simulation/analysis using a sex-averaged map									Simulation/analysis using a sex-specific map									
Hypothesis	$H_{0,a}$	$H_{0,b}$					H_1			$H_{0,a}$	$H_{0,b}$					H_1			
Trait model	–	additive				recessive	$I = 0.4$			–	additive			recessive		$I = 0.4$			
Map ratio	1:1	1:1		9:1		1:1	1:1			9:1	9:1			9:1		9:1			
$MOBIT_{real}$	0.0228	0.4322		1.4376		0.0634	1.417			0.1193	0.4249			0.0382		1.6794			
	Quantiles									Quantiles									
Method	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*	*bfnm*	*perm*
50%	0.2215	0.3553	0.1594	0.1654	0.1363	0.1914	0.0933	0.0915	0.1491	0.1467	0.2046	0.2666	0.1284	0.1647	0.0876	0.1045	0.1747	0.1547		
90%	0.9324	1.0421	0.8125	0.81	0.7461	1.0084	0.6228	0.5731	0.7923	0.9019	0.8372	0.8534	0.7141	0.8175	0.6308	0.6208	0.8543	0.8243		
95%	1.2268	1.3363	1.1142	1.1013	1.0282	1.3339	0.8707	0.7931	1.0998	1.252	1.1055	1.0695	1.0062	1.1086	0.9134	0.8857	1.1323	1.1631		
99%	1.8794	1.9248	1.7467	1.8	1.7173	1.9781	1.4697	1.3598	1.776	2.0126	1.7386	1.5886	1.6469	1.8255	1.5243	1.53	1.8259	1.9375		
p value	0.8069	0.8797	0.2478	0.2492	0.0196	0.039	0.5677	0.5747	0.0241	0.0353	0.6203	0.7068	0.2095	0.2599	0.6269	0.6895	0.0144	0.0171		

$MOBIT_{real}$ corresponds to the MOBIT of the picked dataset from the main simulation study, which served as the real dataset for the simulations/permutations. The empiric p value is the proportion of MOBIT values of the simulation/permutation distribution that are equal to or larger than $MOBIT_{real}$. *bfnm*: simulation method to obtain an empiric p value for the MOBIT using genotype simulations based on the best-fitting nonimprinting model of the real dataset analysis; *perm*: permutation method to obtain an empiric p value for the MOBIT based on the permutation of parental sexes. For more details see Appendix Table 8.

C.1.1 Permutation distribution for ASPs

As can be seen from the results of the sex-averaged ASP analyses in Appendix Figure 10, the permutation procedure generated replicates with an expected equal allele-sharing between the parental sexes on the possible triangle. Points in terms of randomly unequal parental allele-sharing, which correspond to the permutation replicates, were stretched out on a more or less straight line perpendicular to the possible triangle, except for the point of no linkage, where these points were lined up on the outer edges of the tetrahedron. It is of note that the best-fitting nonimprinting model for every replicate always corresponded to the same point in terms of allele-sharing as for the real dataset, which illustrates that the permutation procedure in fact generates replicates under the conditional null hypothesis mentioned above. The same held also true for the permutations using the sex-specific map in Appendix Figure 11, however, the expected point in terms of allele-sharing was shifted into the left half-tetrahedron according to the underlying map ratio. In the case of the point of no linkage, replicates were lined up on either edge of the tetrahedron according to the map ratio. As can be deduced from the points in terms of allele-sharing of replicates for the dataset originally simulated under H_1 (black points, Appendix Figure 11), the permutation procedure reflects peculiarities and boundaries of the parameter space as well as effects due to sex-specific recombination fractions. With regard to the scenario in which a truly underlying 9:1 map ratio is not taken into account in the analysis (red points, Appendix Figure 10), the corresponding quantiles were comparable to those of the 1:1 scenarios, indicating that confounding cannot be prevented using such a permutation procedure. The 95% quantiles of the $H_{0,\,b}$ and H_1, scenarios using a sex-averaged map for method *perm* (see Appendix Table 9, top) were all similar to each other. Furthermore, the corresponding quantiles of the sex-specific analyses were quite similar to the sex-averaged ones, which shows that the MOBIT permutation distribution is less affected by maximization curves, according to the sex-specific map ratio, than the *bfnm* method (see Appendix Table 9, top). The latter effect, however, may cause substantial differences in quantiles and p values between the two methods, and hence also regarding the conclusion whether imprinting is present or not.

C.1.2 Permutation distribution of 3-G pedigrees

The results of the MOBIT permutation distribution for 3-G pedigrees can be found in Appendix Table 9, bottom. The properties of the quantiles were comparable to those for ASPs and seemed to reflect peculiarities of the 3-G pedigree parameter space as can be deduced from the different quantiles for additive and recessive trait models. Similar to ASPs, quantiles for the method *perm* using a sex-specific map in the analysis were also comparable to those obtained from the corresponding analyses using a sex-averaged map. Unlike for ASPs, the quantiles for the *perm* method differed from those obtained by the *bfnm* method not only when using a sex-specific map, but also when using a sex-averaged map in the analysis.

References

Abecasis, G. R., S. S. Cherny, W. O. Cookson and L. R. Cardon (2002): "Merlin–rapid analysis of dense genetic maps using sparse gene flow trees," Nat. Genet., 30, 97–101.

Bain, S. C., B. R. Rowe, A. H. Barnett and J. A. Todd (1994): "Parental origin of diabetes-associated HLA types in sibling pairs with type I diabetes," Diabetes, 43, 1462–1468.

Brugger, M., S. Rospleszcz and K. Strauch (2016): "Estimation of trait model parameters in a MOD score linkage analysis," Hum. Hered., 82, 103–139.

Brugger, M. and K. Strauch (2014): "Fast linkage analysis with MOD scores using algebraic calculation," Hum. Hered., 78, 179–194.

Christensen, U., S. Møller-Larsen, M. Nyegaard, A. Haagerup, A. Hedemand, C. Brasch-Andersen, T. A. Kruse, T. J. Corydon, M. Deleuran and A. D. Børglum (2009): "Linkage of atopic dermatitis to chromosomes 4q22, 3p24 and 3q21," Hum. Genet., 126, 549–557.

Churchill, G. A. and R. W. Doerges (2008): "Naive application of permutation testing leads to inflated type I error rates," Genetics, 178, 609–610.

Clerget-Darpoux, F., C. Bonaïti-Pellié and J. Hochez (1986): "Effects of misspecifying genetic parameters in lod score analysis," Biometrics, 42, 393–399.

Davies, W., A. R. Isles and L. S. Wilkinson (2005): "Imprinted gene expression in the brain," Neurosci. Biobehav. Rev., 29, 421–430.

Daw, E. W., E. A. Thompson and E. M. Wijsman (2000): "Bias in multipoint linkage analysis arising from map misspecification," Genet. Epidemiol., 19, 366–380.

Dib, C., S. Fauré, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette and J. Weissenbach (1996): "A comprehensive genetic map of the human genome based on 5,264 microsatellites," Nature, 380, 152–154.

Dietter, J., M. Mattheisen, R. Fürst, F. Rüschendorf, T. F. Wienker and K. Strauch (2007): "Linkage analysis using sex-specific recombination fractions with GENEHUNTER-MODSCORE," Bioinformatics, 23, 64–70.

Dong, C., W.-D. Li, F. Geller, L. Lei, D. Li, O. Y. Gorlova, J. Hebebrand, C. I. Amos, R. D. Nicholls and R. A. Price (2005): "Possible genomic imprinting of three human obesity-related genetic loci," Am. J. Hum. Genet., 76, 427–437.
Elston, R. C. (1989): "Man bites dog? The validity of maximizing lod scores to determine mode of inheritance," Am. J. Med. Genet., 34, 487–488.
Falls, J. G., D. J. Pulford, A. A. Wylie and R. L. Jirtle (1999): "Genomic imprinting: implications for human disease," Am. J. Pathol., 154, 635–647.
Feng, R. and H. Zhang (2008): "A genomic imprinting test for ordinal traits in pedigree data," Genet. Epidemiol., 32, 132–142.
Fingerlin, T. E., G. R. Abecasis and M. Boehnke (2006): "Using sex-averaged genetic maps in multipoint linkage analysis when identity-by-descent status is incompletely known," Genet. Epidemiol., 30, 384–396.
Fishelson, M., and D. Geiger (2002): "Exact genetic linkage computations for general pedigrees," Bioinformatics, 18(Suppl 1), S189–S198.
Flaquer, A., C. Baumbach, E. Piñero, F. García Algas, M. A. de la Fuente Sanchez, J. Rosell, J. Toquero, L. Alonso-Pulpon, P. Garcia-Pavia, K. Strauch and D. Heine-Suñer (2013): "Genome-wide linkage analysis of congenital heart defects using MOD score analysis identifies two novel loci," BMC Genet., 14, 44.
Flaquer, A., C. Fischer and T. F. Wienker (2009): "A new sex-specific genetic map of the human pseudoautosomal regions (PAR1 and PAR2)," Hum. Hered., 68, 192–200.
Flaquer, A. and K. Strauch (2012): "A comparison of different linkage statistics in small to moderate sized pedigrees with complex diseases," BMC Res. Notes, 5, 411.
Greenberg, D. A., M. Durner, M. Keddache, S. Shinnar, S. R. Resor, S. L. Moshe, D. Rosenbaum, J. Cohen, C. Harden, H. Kang, S. Wallace, D. Luciano, K. Ballaban-Gil, L. Tomasini, G. Zhou, I. Klotz and E. Dicker (2000): "Reproducibility and complications in gene searches: linkage on chromosome 6, heterogeneity, association, and maternal inheritance in juvenile myoclonic epilepsy," Am. J. Hum. Genet., 66, 508–516.
Greenberg, D. A., M. C. Monti, B. Feenstra, J. Zhang and S. E. Hodge (2010): "The essence of linkage-based imprinting detection: comparing power, type 1 error, and the effects of confounders in two different analysis approaches," Ann. Hum. Genet., 74, 248–262.
Gudbjartsson, D. F., K. Jonasson, M. L. Frigge and A. Kong (2000): "Allegro, a new computer program for multipoint linkage analysis," Nat. Genet., 25, 12–13.
Gudbjartsson, D. F., T. Thorvaldsson, A. Kong, G. Gunnarsson and A. Ingolfsdottir (2005): "Allegro version 2," Nat. Genet., 37, 1015–1016.
Hager, R., J. M. Cheverud and J. B. Wolf (2008): "Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting," Genetics, 178, 1755–1762.
Han, M., Y.-Q. Hu and S. Lin (2013): "Joint detection of association, imprinting and maternal effects using all children and their parents," Eur. J. Hum. Genet., 21, 1449–1456.
Hanson, R. L., S. Kobes, R. S. Lindsay and W. C. Knowler (2001): "Assessment of parent-of-origin effects in linkage analysis of quantitative traits," Am. J. Hum. Genet., 68, 951–962.
Hoggart, C. J., G. Venturini, M. Mangino, F. Gomez, G. Ascari, J. H. Zhao, A. Teumer, T. W. Winkler, N. Tšernikova, J. Luan, E. Mihailov, G. B. Ehret, W. Zhang, D. Lamparter, T. O. Esko, A. Macé, S. Rüeger, P.-Y. Bochud, M. Barcella, Y. Dauvilliers, B. Benyamin, D. M. Evans, C. Hayward, M. F. Lopez, L. Franke, A. Russo, I. M. Heid, E. Salvi, S. Vendantam, D. E. Arking, E. Boerwinkle, J. C. Chambers, G. Fiorito, H. Grallert, S. Guarrera, G. Homuth, J. E. Huffman, D. Porteous, D. Moradpour, A. Iranzo, J. Hebebrand, J. P. Kemp, G. J. Lammers, V. Aubert, M. H. Heim, N. G. Martin, G. W. Montgomery, R. Peraita-Adrados, J. Santamaria, F. Negro, C. O. Schmidt, R. A. Scott, T. D. Spector, K. Strauch, H. Völzke, N. J. Wareham, W. Yuan, J. T. Bell, A. Chakravarti, J. S. Kooner, A. Peters, G. Matullo, H. Wallaschofski, J. B. Whitfield, F. Paccaud, P. Vollenweider, S. Bergmann, J. S. Beckmann, M. Tafti, N. D. Hastie, D. Cusi, M. Bochud, T. M. Frayling, A. Metspalu, M.-R. Jarvelin, A. Scherag, G. D. Smith, I. B. Borecki, V. Rousson, J. N. Hirschhorn, C. Rivolta, R. J. F. Loos and Z. Kutalik (2014): "Novel approach identifies SNPs in SLC2a10 and KCNK9 with evidence for parent-of-origin effect on body mass index," PLoS Genet., 10, e1004508.
Holmans, P. (1993): "Asymptotic properties of affected-sib-pair linkage analysis," Am. J. Hum. Genet., 52, 362–374.
Knapp, M. (2005): "A note on linkage analysis with affected sib triplets," Hum. Hered., 59, 21–25.
Knapp, M., S. A. Seuchter and M. P. Baur (1994): "Linkage analysis in nuclear families. 2: Relationship between affected sib-pair tests and lod score analysis," Hum. Hered., 44, 44–51.
Knapp, M. and K. Strauch (2004): "Affected-sib-pair test for linkage based on constraints for identical-by-descent distributions corresponding to disease models with imprinting," Genet. Epidemiol., 26, 273–285.
Kong, A. and N. J. Cox (1997): "Allele-sharing models: LOD scores and accurate linkage tests," Am. J. Hum. Genet., 61, 1179–1188.
Kruglyak, L., M. J. Daly, M. P. Reeve-Daly and E. S. Lander (1996): "Parametric and nonparametric linkage analysis: a unified multipoint approach," Am. J. Hum. Genet., 58, 1347–1363.
Kruse, L. V., M. Nyegaard, U. Christensen, S. Møller-Larsen, A. Haagerup, M. Deleuran, L. G. Hansen, S. K. Venø, D. Goossens, J. Del-Favero and A. D. Børglum (2012): "A genome-wide search for linkage to allergic rhinitis in Danish sib-pair families," Eur. J. Hum. Genet., 20, 965–972.
Kurz, T., J. Altmueller, K. Strauch, F. Rüschendorf, A. Heinzmann, M. F. Moffatt, W.O.C. M. Cookson, F. Inacio, P. Nürnberg, H. H. Stassen and K. A. Deichmann (2005): "A genome-wide screen on the genetics of atopy in a multiethnic European population reveals a major atopy locus on chromosome 3q21.3," Allergy, 60, 192–199.
Kurz, T., K. Strauch, A. Heinzmann, S. Braun, M. Jung, F. Rüschendorf, M. F. Moffatt, W. O. Cookson, F. Inacio, A. Ruffilli, G. Nordskov-Hansen, G. Peltre, J. Forster, J. Kuehr, A. Reis, T. F. Wienker and K. A. Deichmann (2000): "A European study on the genetics of mite sensitization," J. Allergy Clin. Immunol., 106, 925–932.
Lander, E. S. and P. Green (1987): Construction of multilocus genetic linkage maps in humans," Proc. Natl. Acad. Sci. U.S.A., 84, 2363–2367.
Lemire, M. (2005): "A simple nonparametric multipoint procedure to test for linkage through mothers or fathers as well as imprinting effects in the presence of linkage," BMC Genet., 6(Suppl 1), S159.
Lemire, M. (2006): "SUP: an extension to SLINK to allow a larger number of marker loci to be simulated in pedigrees conditional on trait values," BMC Genet., 7, 40.
Lewis, A. and W. Reik (2006): "How imprinting centres work," Cytogenet. Genome Res., 113, 81–89.

Liu, X.-Q., C. Greenwood, K.-S. Wang and A. Paterson (2005): "A genome scan for parent-of-origin linkage effects in alcoholism," BMC Genet., 6(Suppl 1), S160.
Maeda, N. and Y. Hayashizaki (2006): "Genome-wide survey of imprinted genes," Cytogenet. Genome Res., 113, 144–152.
Matise, T. C., F. Chen, W. Chen, F. M. D. L. Vega, M. Hansen, C. He, F. C. L. Hyland, G. C. Kennedy, X. Kong, S. S. Murray, J. S. Ziegle, W. C. L. Stewart and S. Buyske (2007): "A second-generation combined linkage physical map of the human genome," Genome Res., 17, 1783–1786.
Mattheisen, M., J. Dietter, M. Knapp, M. P. Baur and K. Strauch (2008): "Inferential testing for linkage with GENEHUNTER-MODSCORE: the impact of the pedigree structure on the null distribution of multipoint MOD scores," Genet. Epidemiol., 32, 73–83.
Metz, C. W. (1938): "Chromosome behaviour, inheritance and sex determination in *Sciara*," Am. Nat., 72, 485–520.
Moffatt, M. F. and W. O. Cookson (1998): "The genetics of asthma. Maternal effects in atopic disease," Clin. Exp. Allergy, 28(Suppl 1), 56–61; discussion 65–6.
Mukhopadhyay, N. and D. E. Weeks (2003): "Linkage analysis of adult height with parent-of-origin effects in the Framingham Heart Study," BMC Genet., 4(Suppl 1), S76.
Ott, J. (1989): "Computer-simulation methods in human linkage analysis," Proc. Natl. Acad. Sci. U.S.A., 86, 4175–4178.
Pàldi, A., G. Gyapay and J. Jami (1995): "Imprinted chromosomal regions of the human genome display sex-specific meiotic recombination frequencies," Curr. Biol., 5, 1030–1035..
Paterson, A. D., D. M. Naimark and A. Petronis (1999): The analysis of parental origin of alleles may detect susceptibility loci for complex disorders," Hum. Hered., 49, 197–204.
Risch, N. (1984): "Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes," Am. J. Hum. Genet., 36, 363–386.
Schäffer, A. A., M. Lemire, J. Ott, G. M. Lathrop and D. E. Weeks (2011): "Coordinated conditional simulation with SLINK and SUP of many markers linked or associated to a trait in large pedigrees," Hum. Hered., 71, 126–134.
Schumacher, J., R. Kaneva, R. A. Jamra, G. O. Diaz, S. Ohlraun, V. Milanova, Y.-A. Lee, F. Rivas, F. Mayoral, R. Fuerst, A. Flaquer, C. Windemuth, E. Gay, S. Sanz, M. J. González, S. Gil, F. Cabaleiro, F. del Rio, F. Perez, J. Haro, C. Kostov, V. Chorbov, A. Nikolova-Hill, V. Stoyanova, G. Onchev, I. Kremensky, K. Strauch, T. G. Schulze, P. Nürnberg, W. Gaebel, A. Klimke, G. Auburger, T. F. Wienker, L. Kalaydjieva, P. Propping, S. Cichon, A. Jablensky, M. Rietschel and M. M. Nöthen (2005): "Genomewide scan and fine-mapping linkage studies in four European samples with bipolar affective disorder suggest a new susceptibility locus on chromosome 1p35-p36 and provides further evidence of loci on chromosome 4q31 and 6q24," Am. J. Hum. Genet., 77, 1102–1111.
Self, S. and K. Liang (1987): "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions," J. Am. Stat. Assoc., 82, 605–610.
Shete, S. and X. Zhou (2005): "Parametric approach to genomic imprinting analysis with applications to Angelman's syndrome," Hum. Hered., 59, 26–33.
Shete, S., X. Zhou and C. I. Amos (2003): "Genomic imprinting and linkage test for quantitative-trait loci in extended pedigrees," Am. J. Hum. Genet., 73, 933–938.
Shute, N. C. and W. J. Ewens (1988): "A resolution of the ascertainment sampling problem. III. Pedigrees," Am. J. Hum. Genet., 43, 387–395.
Sieberts, S. K. and D. F. Gudbjartsson (2005): "Sex-specific maps and consequences for linkage mapping," In Lynn Jorde, Peter Little, Mike Dunn, and Shankar Subramaniam, (Eds.), Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. John Wiley & Sons, Chichester, UK.
Smalley, S. L. (1993): "Sex-specific recombination frequencies: a consequence of imprinting?," Am. J. Hum. Genet., 52, 210–212.
Solter, D. (2006): "Imprinting today: end of the beginning or beginning of the end?" Cytogenet. Genome Res., 113, 12–16.
Spencer, H. G. (2009): "Effects of genomic imprinting on quantitative traits," Genetica, 136, 285–293.
Stine, O. C., J. Xu, R. Koskela, F. J. McMahon, M. Gschwend, C. Friddle, C. D. Clark, M. G. McInnis, S. G. Simpson, T. S. Breschel, E. Vishio, K. Riskin, H. Feilotter, E. Chen, S. Shen, S. Folstein, D. A. Meyers, D. Botstein, T. G. Marr and J. R. DePaulo (1995): "Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect," Am. J. Hum. Genet., 57, 1384–1394.
Strauch, K. (2003): "Parametric linkage analysis with automatic optimization of the disease model parameters," Am. J. Hum. Genet., 73, A2624.
Strauch, K. (2005): "Gene mapping, imprinting and epigenetics," In Lynn Jorde, Peter Little, Mike Dunn, and Shankar Subramaniam, (Eds.), Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. John Wiley & Sons, Chichester, UK.
Strauch, K. (2007): "MOD-score analysis with simple pedigrees: an overview of likelihood-based linkage methods," Hum. Hered., 64, 192–202.
Strauch, K., R. Fimmers, T. Kurz, K. A. Deichmann, T. F. Wienker and M. P. Baur (2000a): "Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization," Am. J. Hum. Genet., 66, 1945–1957.
Strauch, K., R. Fimmers, T. F. Wienker, M. P. Baur, S. Cichon, P. Propping and M. M. Nöthen (2000b): "Strauch et al reply," Mol. Psychiatry, 5, 126–127.
Suarez, B. K., J. Rice and T. Reich (1978): "The generalized sib pair IBD distribution: its use in the detection of linkage," Ann. Hum. Genet., 42, 87–94.
Vincent, Q., A. Alcaïs, A. Alter, E. Schurr and L. Abel (2006): "Quantifying genomic imprinting in the presence of linkage," Biometrics, 62, 1071–1080.
Walter, J. and M. Paulsen (2003): "Imprinting and disease," Semin. Cell Dev. Biol., 14, 101–110.
Weeks, D. E. (2010): "powerpkg: Power analyses for the affected sib pair and the TDT design," R package version 1.3. https://CRAN.R-project.org/package=powerpkg.
Weeks, D. E., T. Lehner, E. Squires-Wheeler, C. Kaufmann and J. Ott (1990): "Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis," Genet. Epidemiol., 7, 237–243.
Whittaker, J. C., N. Gharani, P. Hindmarsh and M. I. McCarthy (2003): "Estimation and testing of parent-of-origin effects for quantitative traits," Am. J. Hum. Genet., 72, 1035–1039.

Wilkins-Haug, L. (2009): "Epigenetics and assisted reproduction," Curr. Opin. Obstet. Gynecol., 21, 201–206.
Wilks, S. S. (1938): "The large-sample distribution of the likelihood ratio for testing composite hypotheses," Ann. Math. Stat., 9, 60–62.
Wu, C.-C., S. Shete and C. I. Amos (2005): "Linkage analysis of affected sib pairs allowing for parent-of-origin effects," Ann. Hum. Genet., 69, 113–126.

Danksagung

Die vorliegende Arbeit wäre ohne die Liebe, Opferbereitschaft und Unterstützung mehrerer, mir über Jahre wohlgesonnener Menschen nicht möglich gewesen. Allen diesen Menschen sei an dieser Stelle für ihren Anteil an dieser Arbeit gedankt! Weil es sich so gehört, und ich auch denke, dass sie es sich verdient haben, möchte ich mich bei einigen Menschen im Folgenden persönlich für ihr Engagement bedanken.

Natürlich bin ich zuallererst meinem Doktorvater Prof. Dr. Konstantin Strauch zu grenzenlosem Dank verpflichtet. Als Sie mich im Frühjahr 2007 in meiner Wohnung in der Kugelgasse 17 in Marburg anriefen und mir eine Stelle als studentische Hilfskraft anboten, konnte ich nicht ahnen, dass wir einen so langen und erfolgreichen, gemeinsamen Weg gehen würden. Ohne Ihren Mut, Ihre Weitsicht und Ihr Vertrauen in mich wäre diese Dissertation nie zustande gekommen. Wissenschaftlich zu arbeiten und zu denken – das habe ich alles von Ihnen gelernt.

Meinen Eltern danke ich für alles, was ich in meinem bisherigen Leben erreicht habe. Euch ist diese Arbeit in Dankbarkeit gewidmet. Ich liebe Euch!

Ich danke meiner Frau Anja und meinen Kindern Paula und Peter, die aufgrund dieser Arbeit viel zu entbehren hatten und mich trotzdem immer wieder antrieben, den Weg bis hierher weiterzugehen. Ich liebe Euch!

Bei Clemens Baumbach bedanke ich mich für eine langjährige Freundschaft und Verbundenheit sowie für zahlreiche schlaflose Nächte, in denen wir ausgiebig über wissenschaftliche Fragen, schlechte und gute Programmierstile sowie die neuesten Tennisergebnisse diskutieren konnten.

Meinem langjährigen Studienfreund und Wanderbruder Kai Siebert alias Django Rosetti danke ich für seine Freundschaft und die vielen, bereichernden Erkenntnisse, zu denen wir über viele Jahre hinweg auf unseren Wallfahrten gekommen sind. Ich freue mich schon auf unsere nächste Tour!