Aus dem Institut für Medizinische Biometrie und Epidemiologie
Geschäftsführender Direktor: Prof. Dr. H. Schäfer
des Fachbereichs Medizin der Philipps-Universität Marburg

Methoden und Algorithmen der Kopplungsanalyse bei quantitativen Phänotypen

Inaugural-Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften

dem Fachbereich Medizin der Philipps-Universität Marburg
vorgelegt von

Thomas Künzel

aus München

Marburg, 2012

Angenommen vom Fachbereich Medizin der Philipps-Universität Marburg
am: 13. Juli 2012

Gedruckt mit Genehmigung des Fachbereichs.

Dekan: Prof. Dr. med. Matthias Rothmund

Referent: Prof. Dr. rer. nat. Konstantin Strauch

Korreferent: Prof. Dr. med. Karl-Heinz Grzeschik

Einleitung

Diese Arbeit ist auf dem Gebiet der genetischen Epidemiologie bzw. statistischen Genetik angesiedelt. Gegenstand dieser beiden Forschungsgebiete ist u. a. die Identifizierung und Charakterisierung von Genen, die ursächlich an der Entstehung von Krankheiten beim Menschen beteiligt sind. Zu diesem Zweck werden polymorphe genetische Marker typisiert, häufig in dichter Abdeckung aller Chromosomen (sogenannter Genom-Scan). Eines der statistischen Verfahren der genetischen Epidemiologie ist die Kopplungsanalyse. Mittels der Kopplungsanalyse prüft man, ob das Vererbungsmuster der Markerallele innerhalb eines Stammbaums mit jenem der Krankheit übereinstimmt. Ist dies der Fall, schließt man auf die Existenz eines krankheits(mit)verursachenden Gens in der Nähe des betreffenden genetischen Markers. Bei genetisch komplexen Krankheiten ist der Effekt eines einzelnen Gens gering, dementsprechend schwierig gestaltet sich die Lokalisierung. Für eine erfolgreiche Genkartierung ist es deshalb wichtig, den Erbgang bzw. das Krankheitsmodell des untersuchten Phänotyps möglichst realistisch zu modellieren.

Diese Arbeit befasst sich mit der Weiterentwicklung von Methoden der Kopplungsanalyse für quantitative Phänotypen, beispielsweise Blutdruck, Serumkonzentration bestimmter Stoffe oder Bildgebungsparameter. Quantitative Phänotypen beeinflussen häufig als sogenannte intermediäre Phänotypen das Auftreten von genetisch komplexen Krankheiten. Dies ist z. B. im Rahmen eines Treshold-Modells denkbar, bei dem alle Personen betroffen sind, deren quantitativer Phänotyp einen bestimmten Wert überschreitet. Bei der bereits verfügbaren MOD-Score-Analyse für dichotome Merkmale wird die Prüfgröße des Tests auf Kopplung, der LOD-Score, über die Parameter des Krankheitsmodells maximiert, was zu einer höheren Power als eine Analyse unter einem fest vorgegebenen Krankheitsmodell führen kann. Ein solches Verfahren wird nun hier für quantitative Phänotypen entwickelt und die mathematischen Methoden der Maximierung vorgestellt. Genomisches Imprinting, Erbgänge mit oder ohne Dominanzeffekte sowie gleiche bzw. unterschiedliche Restvarianzen

der genotypspezifischen Verteilungen werden hierbei ebenfalls berücksichtigt. Anhand von realen als auch simulierten Daten wird die Methode dann im Vergleich mit bestehenden Verfahren, der Varianzkomponentenanalyse (VCA) und der Haseman-Elston-Regression, validiert. Das Verfahren wurde in das Software Paket GENEHUNTER implementiert, die Erweiterung für quantitative Phänotypen trägt den Namen GENEHUNTER-QMOD.

Diese methodischen Entwicklungen liefern einen wichtigen Beitrag zur besseren Lokalisierbarkeit von Genen, die quantitativen Phänotypen unterliegen und dadurch komplexe Krankheiten verursachen oder wenigstens zu ihrer Entstehung beitragen. Weiterhin helfen die durch die Modellierung der quantitativen Phänotypen gewonnenen Informationen, die zugrundeliegenden stoffwechselphysiologischen Mechanismen besser zu verstehen.

Inhaltsverzeichnis

Tabellenverzeichnis

Bildverzeichnis

1 Kopplungsanalyse

Krankheiten beim Menschen werden zu einem großen Teil durch genetische Varianten beeinflusst oder verursacht. Um den Krankheitsmechanismus zu verstehen und um Patienten ursächlich behandeln zu können, ist ein erster Schritt, die genetische Variante im menschlichen Genom zu lokalisieren. Ein wichtiges Hilfsmittel für dieses Ziel ist die *Kopplungsanalyse*. Die Kopplungsanalyse nutzt die Tatsache, dass Allele an Genorten oder Loci, die physisch im Genom nahe beieinander liegen, auch häufig miteinander vererbt werden. Allele an Genorten, die weiter voneinander entfernt liegen, werden über die Generationen weit häufiger getrennt und separat weitervererbt.

1.1 Vererbung und Rekombination

Wie genau funktioniert Vererbung auf dem menschlichen Genom, und wann werden Allele zusammen bzw. getrennt weitervererbt? Dieser Frage gehen wir in diesem Kapitel nach.

Menschliche Körperzellen besitzen einen doppelten Chromosomensatz von je 23 Chromosomen. Man spricht auch von einem diploiden Genom, was bedeutet, dass alle 23 Chromosomen zweimal in jeder Körperzelle existieren. 22 dieser Chromosomen sind Autosomen. Je zwei Chromosomen der gleichen Sorte besitzen Erbinformationen für die selben Eigenschaften. Eine Ausnahme bildet jedoch das 23. Chromosomenpaar, die Geschlechtschromosomen oder Gonosomen. In weiblichen Zellen finden sich hier zwei (homologe) X-Chromosomen, in männlichen Zellen jedoch ein X- und ein Y-Chromosom. Letzteres ist bedeutend kürzer als das X-Chromosom, somit sind diese beiden Chromosomen nicht homolog.

Um Erbgut weiterzugeben, bilden sich im Menschen Geschlechtszellen, auch Gameten genannt, aus. Diese enthalten von jedem Chromosomenpaar nur eines der beiden Chromosomen. Man nennt diesen Chromosomensatz auch haploid. Zwei haploide Gameten, eine männliche und eine weibliche, verschmelzen zu einer Zygote. Diese

enthält nun beide haploide Chromosomensätze der Gameten und ist somit wieder diploid. Aus ihr entsteht dann der Nachkomme, der nun die Hälfte seines Erbguts von der männlichen Gamete und die andere Hälfte von der weiblichen Gamete erhalten hat. Der biologische Vorgang, bei dem Geschlechtszellen entstehen, heißt Meiose. Während dieses Prozesses ordnen sich die homologen Chromosomen der Länge nach nebeneinander an, siehe Bild 1.1 links. Bevor sich die homologen Chromosomen auftrennen und auf die Gameten aufgeteilt werden, überkreuzen sich die Chromosomen an zufälligen Stellen, siehe Bild 1.1 Mitte. Diese nennt man Chiasmata. Die Chromosomen trennen sich nun an dieser Stelle auf und verbinden sich mit dem jeweils anderen Chromosom. Hier hat also ein Austausch von Erbgut stattgefunden, siehe Bild 1.1 rechts. Diesen Vorgang nennt man Crossover. Wie ebenfalls in Bild 1.1 zu sehen, sind mehrere Crossover auf demselben Chromosomenpaar möglich. Es kann also sein, dass Erbgut "zurückgetauscht" wird und sich trotz Crossover auf dem selben Chromosom befindet wie vorher, z.B. Locus A und C in Bild 1.1. Sowohl die Allele A1 und C1 als auch die Allele A2 und C2 befinden sich vor und nach den Crossovern auf dem selben Chromosom. Betrachtet man jedoch zwei Loci auf dem Chromosom, zwischen denen eine ungerade Anzahl an Crossovern stattgefunden hat, so wurden die Allele an diesen Loci durch Crossover vertauscht, z.B. Locus A und B in Bild 1.1. Ist dies geschehen, spricht man von Rekombination. Dieser Begriff ist für die Kopplungsanalyse von fundamentaler Bedeutung, wie wir gleich sehen werden. Nun ist nämlich ersichtlich, warum nahe beieinander liegende Allele häufig gemeinsam vererbt werden und weit entfernte nicht: Je größer der Abstand zweier Allele auf einem Chromosom, desto höher ist die Wahrscheinlichkeit für ein oder mehrere Chiasmata zwischen den beiden Allelen. Je weiter sie voneinander entfernt liegen, desto höher ist also auch die Wahrscheinlichkeit, dass sie rekombinieren und damit getrennt vererbt werden. Bei benachbarten Allelen geschieht dies so gut wie nie, sie werden fast immer gemeinsam vererbt. Ein Maß für den Abstand zweier Loci auf einem Chromosom ist die sog. *Rekombinationsfrequenz* θ. Sie gibt die Wahrscheinlichkeit an, mit der die beiden Loci während einer Meiose rekombinieren. Es gilt $\theta \in \left[0, \frac{1}{2}\right]$. Dabei bedeutet $\theta = 0$, dass beide Loci direkt nebeneinander liegen, und vollständige Kopplung vorliegt. $\theta = \frac{1}{2}$ bedeutet, dass beide Loci weit voneinander entfernt bzw. auf verschiedenen Chromosomen liegen.

Ein weiteres Maß für die Entfernung zweier Loci ist der *genetische Abstand* x. Er wird üblicherweise in centiMorgan (cM) angegeben. Ein Abstand zweier Loci von

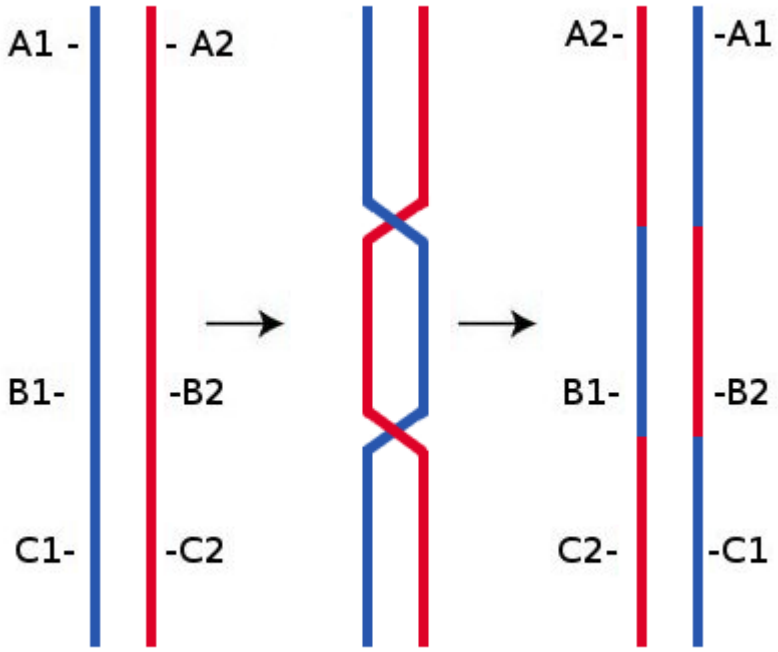

Bild 1.1: Crossover

$x = 1cM$ bedeutet, dass in dem Abschnitt zwischen den Loci mit Wahrscheinlichkeit 0,01 ein Chiasma stattfindet. Dies ist jedoch nicht mit einer Rekombinationsfrequenz von $\theta = 0{,}01$ gleichzusetzen, da mehrere Chiasmata/Crossover auftreten können, und Rekombination wie erwähnt nur bei einer ungeraden Anzahl von Crossovern stattfindet. Der Zusammenhang zwischen Rekombinationsfrequenz θ und genetischem Abstand x lässt sich durch eine sogenannte *Mapping-Funktion* beschreiben. Welche Mapping-Funktion den Zusammenhang korrekt beschreibt, hängt davon ab, ob Interferenz zwischen den einzelnen Chiasmata stattfindet. Dies bedeutet, dass ein Chiasma die Wahrscheinlichkeit für ein weiteres Chiasma in seiner Umgebung beeinflusst. Meistens sinkt aus vor allem räumlichen Gründen die Wahrscheinlichkeit für ein weiteres Chiasma in der Nähe eines ersten. Ohne Interferenz gilt die Mapping-Funktion nach Haldane:

$$\begin{aligned} x &= -\frac{1}{2}\ln(1-2\theta) \\ \theta &= \frac{1}{2}\left(1-e^{-2|x|}\right). \end{aligned} \tag{1.1.1}$$

Mit Interferenz ist der Zusammenhang anders zu formulieren. Ein gewisses Maß an positiver Interferenz (Wahrscheinlichkeit für ein zweites Chiasma neben einem ersten sinkt) berücksichtigt z.B. die Mapping-Funktion von Kosambi:

$$\begin{aligned} x &= \frac{1}{2}\tanh^{-1}(2\theta) &&= \frac{1}{4}\ln\frac{1+2\theta}{1-2\theta} \\ \theta &= \frac{1}{2}\tanh(2x) &&= \frac{1}{2}\frac{e^{4x}-1}{e^{4x}+1}. \end{aligned} \tag{1.1.2}$$

1.2 Der Test auf Kopplung

Das Grundprinzip der Kopplungsanalyse baut auf der gemeinsamen bzw. durch Rekombination getrennten Vererbung von genetischen Polymorphismen auf. Hierbei handelt es sich um einen kleinen Ausschnitt aus dem Genom, dessen Position man kennt und der bei verschiedenen Personen auch in verschiedenen Varianten vorliegen kann. Einen solchen Polymorphismus bezeichnet man auch als Marker. Im Rahmen der Kartierung krankheitsverursachender Gene wird bei der Kopplungsanalyse überprüft, ob ein Marker, dessen Position bekannt ist, überzufällig häufig mit der untersuchten Krankheit vererbt wird. Man spricht von Kosegregation. Ist dies der Fall, so kann man schließen, dass sich der Krankheitslocus, der für die Entstehung der Krankheit verantwortlich ist, auf demselben Chromosom und nahe dem Marker befindet. Anderenfalls wären Marker und Krankheit über die Generationen hinweg irgendwann einmal durch Rekombination aufgetrennt und separat weitervererbt worden, oder liegen von vornherein auf getrennten Chromosomen, was immer unabhängige Vererbung zur Folge hat. Um Kosegregation zu erkennen, ist es notwendig, dass bei den Markerloci für die Personen, anhand derer die Vererbung untersucht wird, verschiedene Ausprägungen des Markers vorliegen. Nur so kann Rekombination auch erkannt werden. Wäre z.B. in Bild 1.1 A1 = A2, so lägen B1 und A1 sowohl vor als auch nach der Rekombination auf demselben Chromosom, und es wäre nicht ersichtlich, ob sie gemeinsam weitervererbt wurden oder nicht. Einen Marker, der viele verschiedene Ausprägungen besitzt, so dass Kosegregation auch erkannt werden kann, bezeichnet man als hoch polymorph.

Die einfachste Form der Kopplungsanalyse ist die sog. *Zweipunkt-Analyse*. Sie funktioniert nach dem oben beschriebenen Schema: Die Position eines Markerlocus ist bekannt und es wird überprüft, ob die Krankheit bzw. der untersuchte Phänotyp mit diesem kosegregiert. Ist dies der Fall, kann auf die Position des Krankheitslocus geschlossen werden. Eine Erweiterung der Zweipunkt-Analyse ist der Multipoint-Ansatz. Hierbei sind bereits die Positionen einer ganzen Gruppe von Markern bekannt, sie bilden eine sogenannte *marker map*. Damit kennt man auch die Rekombinationsfrequenzen zwischen den einzelnen Markern. Hierbei wird geprüft, ob der Krankheitslocus an die Markergruppe gekoppelt ist bzw. mit einigen der Marker aus der Gruppe kosegregiert.

Neben Zweipunkt- und Multipoint-Analyse gibt es noch ein Kriterium, um Kopplungsanalysemethoden zu differenzieren, nämlich danach, ob sie einen parametrischen oder nicht-parametrischen Ansatz verfolgen. Bei der parametrischen Kopplungsanalyse macht man vor der Analyse Annahmen darüber, nach welchem Schema die Krankheit auftritt bzw. vererbt wird. Ein Beispiel ist eine Krankheit, die durch einen (autosomalen) Locus bestimmt ist, an dem genau zwei allelische Ausprägungen möglich sind. Wir bezeichnen mit "+" das Wildtyp-Allel und mit "m" das mutante Allel oder Krankheitsallel. Macht man keine Unterscheidung zwischen paternal und maternal geerbtem Allel, so ergeben sich am Krankheitslocus die Genotypmöglichkeiten

$$(+,+)\,,\ (m,+)\,,\ (m,m)\,. \tag{1.2.1}$$

Das Krankheitsmodell besteht nun aus der Krankheitsallelfrequenz $p(m)$ sowie den drei Penetranzen

$$f_{+/+},\ f_{m/+},\ f_{m/m}. \tag{1.2.2}$$

Die Krankheitsallelfrequenz bezeichnet die relative Häufigkeit des Allels in der Population. Die Penetranzen geben an, mit welcher Wahrscheinlichkeit eine Person mit dem entsprechenden Genotyp tatsächlich erkrankt. Diese Angabe kann jedoch nur für dichotome Phänotypen/Krankheiten gemacht werden, wie wir später sehen werden. Mit den Penetranzen lassen sich z.B. vollständig dominante Erbleiden

$$f_{+/+} = 0,\ f_{m/+} = 1,\ f_{m/m} = 1 \tag{1.2.3}$$

oder vollständig rezessive Erbleiden

$$f_{+/+} = 0,\ f_{m/+} = 0,\ f_{m/m} = 1 \tag{1.2.4}$$

oder entsprechende Mischformen beschreiben.

Im Gegensatz dazu existieren bei der nicht parametrischen Kopplungsanalyse keine Modellparameter wie Penetranzen oder Krankheitsallelfrequenzen. Man spricht auch von modell-freier Analyse. In diesem Fall wird untersucht, ob Personen mit der Erbkrankheit innerhalb eines Stammbaumes mehr Allele gemeinsam haben, als man unter der Nullhypothese keiner Kopplung erwarten würde. Die nicht parametrische Kopplungsanalyse ist jedoch nicht Gegenstand dieser Arbeit und wird deswegen auch nicht weiter behandelt.

Im Rahmen der Kopplungsanalyse wird ein statistischer Signifikanztest durchgeführt, um zu entscheiden, ob Kopplung vorliegt oder nicht. Bei der parametrischen Kopplungsanalyse verwendet man einen Likelihoodquotiententest. Die Nullhypothese ist hierbei, dass die (bekannten) Marker und der unbekannte Krankheitslocus unabhängig voneinander vererbt werden, d.h. zwischen den beiden Loci gilt $\theta = \frac{1}{2}$ bzw. im Multi-Marker-Fall Position x des Krankheitslocus weit von der Markergruppe entfernt. Die Alternativhypothese hingegen ist, dass Marker und Krankheitslocus kosegregieren, also gemeinsam vererbt werden ($\theta < \frac{1}{2}$ bzw. x in der Nähe der Markergruppe). Die Prüfgröße, anhand derer die Testentscheidung gefällt wird, heißt *Logarithm-of-Odds*-Score (LOD-Score). Dieser ist im parametrischen Fall wie folgt definiert:

$$\mathrm{LOD} := \log_{10}\left(\frac{P\left(\mathrm{Daten}\,|\,\mathrm{Modell},\,\theta\right)}{P\left(\mathrm{Daten}\,|\,\mathrm{Modell},\,\theta=\frac{1}{2}\right)}\right). \tag{1.2.5}$$

Bei den Daten handelt es sich um die beobachteten Krankheitsphänotypen und die Markerallele innerhalb eines oder mehrerer Stammbäume. Zum Modell gehören die Krankheitsallelfrequenz und die Penetranzen (im dichotomen Fall wären dies die drei Penetranzparameter aus (1.2.2)) sowie die Allelfrequenzen am Markergenort (Zweipunkt-Analyse) bzw. den Markergenorten (Multipoint-Analyse). Die Rekombinationsfrequenzen zwischen den einzelnen Markern gehen ebenfalls als Parameter ins Modell ein. Auch die Verwandtschaftsbeziehungen der einzelnen Personen untereinander gehören hierzu.

Der Parameter θ bzw. x ist unbekannt, er bestimmt ja gerade die Position des

Krankheitslocus, die durch die Kopplungsanalyse gefunden werden soll. Deswegen wird der LOD-Score über $\theta \in [0, \frac{1}{2}]$ bzw. x maximiert. Der Wert für θ bzw. x, der den maximalen LOD-Score liefert, ist der Schätzer für die Position des Krankheitslocus.

Die Berechnung des LOD-Scores kann, je nach untersuchtem Datensatz, recht kompliziert sein. Im Kontext der GENEHUNTER-Software ist das Herzstück der Implementierung der sog. *Lander-Green*-Algorithmus (Lander & Green, 1987), der ausführlich

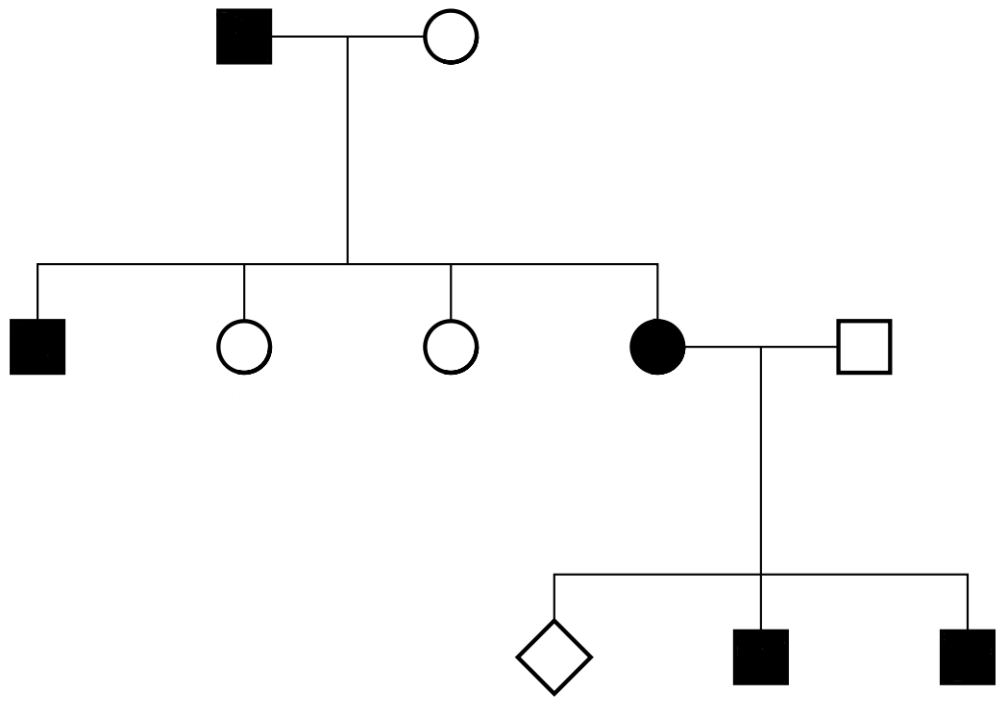

Bild 1.2: Ein exemplarischer Stammbaum: Weibliche Personen werden durch einen Kreis, männliche Personen durch ein Quadrat dargestellt. Ein unbekanntes Geschlecht wird mit einem auf die Spitze gestellten Quadrat symbolisiert. Personen ohne Eltern im Stammbaum bezeichnet man als *Founder*. Bei dichotomen Phänotypen bedeutet ein schwarz ausgefülltes Quadrat (Kreis), dass die entsprechende Person von der untersuchten Krankheit betroffen ist. Nicht ausgefüllte Symbole entsprechen gesunden Personen. Im quantitativen Fall entfällt diese Notation, hier muss der Krankheitsphänotyp numerisch angegeben werden.

von Strauch (2002), S.41 ff beschrieben wurde. Der große Vorteil dieses Algorithmus' ist, dass der Rechenaufwand nur linear mit der Anzahl der verwendeten Marker steigt. Andererseits hängt die Rechenzeit exponentiell von der Anzahl der Personen in den verwendeten Stammbäumen ab. Deswegen lässt sich Software zur Kopplungsanalyse wie GENEHUNTER mit vielen Markern benutzen, allerdings führen Stammbäume mit mehr als 20 effektiven Meiosen zu inakzeptabel hohen Rechenzeiten und Speicheranforderungen auf einem Standard-PC (Quad-Core CPU, 3.0 GHz, 8 GByte RAM).

In vielen Fällen sind die Modellparameter aus (1.2.2) der Krankheit bzw. des Phänotyps nicht bekannt. Da sich eine LOD-Score-Analyse dann nicht durchführen läßt, müssen auch die Modellparameter geschätzt werden. Dies geschieht, in dem die LOD-Score-Funktion (1.2.5) über alle unbekannten Modellparameter mitmaximiert wird. Es wird also eine mehrdimensionale Optimierung durchgeführt. Als Schätzer für die Modellparameter gelten wiederum diejenigen Werte, die den höchsten LOD-Score, jetzt auch MOD-Score genannt, liefern.

Der MOD-Score liefert neben einem Maximum-Likelihood-Schätzer eine Teststatistik für den entsprechenden Signifikanztest. Die Verteilung des MOD-Scores unter der Nullhypothese keiner Kopplung ist im dichotomen Fall für spezielle Stammbaumstrukturen bekannt, siehe Strauch (2007). Damit ist der zugehörige p-Wert leicht erhältlich. In den meisten Fällen ist die Verteilung jedoch unbekannt, und der entsprechende p-Wert wird durch Simulationen durch einen empirischen p-Wert approximiert. Für diesen Fall verfügt GENEHUNTER bereits über entsprechende Simulationsroutinen, die im Rahmen der Version GENEHUNTER-MODSCORE implementiert wurden, siehe Mattheisen *et al.* (2008). Liegt der p-Wert vor, sei es durch die entsprechende Verteilung oder durch Simulationen, kann eine Testentscheidung getroffen werden. Ist der p-Wert genügend klein, wird die Nullhypothese keiner Kopplung verworfen und zugunsten der Alternativhypothese Kopplung entschieden. Die geschätzte Position des Krankheitslocus ist dann $\hat{\theta}$ (oder $\hat{x}$ im Multi-Marker Fall), das den MOD-Score maximiert.

2 Quantitative Phänotypen

In diesem und dem nächsten Kapitel wird das eigentliche Thema dieser Arbeit behandelt, nämlich ein neues Kopplungsanalyse-Verfahren für quantitative Phänotypen. Nach einer kurzen Einleitung über quantitative Phänotypen erfolgt die Entwicklung des in dieser Arbeit verwendeten Modells. Im nächsten Kapitel wird ein Verfahren beschrieben, mit dem der quantitative MOD-Score berechnet, also der LOD-Score am geeignetsten über alle Krankheitsmodellparameter maximiert werden kann. Hierbei handelt es sich um ein mathematisches, mehrdimensionales Optimierungsproblem, das numerisch gelöst wird. Abschließend wird kurz erklärt, wie die Implementierung in GENEHUNTER erfolgte. Die GENEHUNTER-Erweiterung mit dem neuen Verfahren nennt sich GENEHUNTER-QMOD.

2.1 Grundidee

Viele Merkmale oder Krankheiten von Lebewesen lassen sich durch zwei oder mehrere, jedenfalls immer endlich viele Zustände adäquat beschreiben. So ist die Blutgruppe einer Person entweder durch A, B, AB oder 0 gegeben, oder eine Patientin leidet entweder unter Brustkrebs oder sie tut es nicht. Dies erlaubt eine eindeutige Klassifizierung des Patienten in "krank" oder "gesund". Kennt man die Genotypen, die das entsprechende Merkmal beeinflussen, sowie das Vererbungsmuster (dominant, rezessiv, additiv ...), so lassen sich Wahrscheinlichkeiten angeben, mit der die Person, bedingt auf ihren Genotyp, den Zustand "krank" oder "gesund" besitzt.

Andere Merkmale lassen sich auf diese Weise nicht beschreiben. So schwankt z. B. der Gesamtcholesterinspiegel des Menschen, der u. a. durch genetische Faktoren mitbeeinflusst wird, größtenteils zwischen 190 und 280 mg/dl (deutsche Bevölkerung zwischen 35 und 65 Jahren). Er kann also theoretisch jeden beliebigen numerischen Wert in diesem Intervall annehmen, und muss deshalb als kontinuierliche Größe betrachtet werden. Man spricht von einem *quantitativen Phänotypen*. Ma-

thematisch betrachtet lässt sich dieser einfach als Variable $y \in I$ auffassen, wobei $I = [a, b] \subset \mathbb{R}, a, b \in \mathbb{R}$. Eine ausführliche Einleitung zu quantitativen Phänotypen sowie eine Fülle von Verfahren und Methoden finden sich z. B. in Falconer & Mackay (1996) und Camp & Cox (2002).

In dieser Situation entfällt zunächst eine Klassifizierung in "krank" und "gesund". Selbstverständlich ist es möglich, Grenzen zu definieren, außerhalb derer der Cholesterinspiegel als gesundheitsgefährdend anzusehen ist. Eine Person mit einem Cholesterinspiegel oberhalb dieser Grenze könnte man dann als krank, eine Person unterhalb als gesund bezeichnen. Damit hätte man mittels des quantitativen Phänotyps einen dichotomen Phänotyp definiert. In diesem Fall gehen allerdings wichtige Informationen, nämlich die des genauen numerischen Wertes, verloren, die wir im Folgenden nutzen wollen.

Das nächste Problem, das sich stellt, wenn man einen quantitativen Phänotyp untersucht, ist das Nichtvorhandensein von herkömmlichen Penetranzen. Die Wahrscheinlichkeit, dass eine Person aus der Grundgesamtheit einen exakten vorgegebenen Cholesterinspiegel besitzt, ist verschwindend gering. Will man sich ein Modell mit Penetranzen konstruieren, müsste man diese alle auf 0 setzen, was eine weitere sinnvolle Likelihoodberechnung ausschließt. Bleiben wir in unserem mathematischen Modell $y \in I$, so handelt es sich bei y um eine überabzählbare Größe, womit sich eine Zuweisung von Wahrscheinlichkeiten an die Elementarereignisse ebenfalls verbietet. Es ist also ein anderer Ansatz vonnöten.

Zu diesem Zweck interpretieren wir den quantitativen Phänotyp als kontinuierliche Zufallsvariable y mit zugehöriger Wahrscheinlichkeitsdichtefunktion f. Viele Merkmale oder Krankheiten, die man untersucht, sind näherungsweise normalverteilt, deswegen werden auch wir uns in unserem Modell für die Normalverteilung entscheiden. Der untersuchte Phänotyp gehorche also im Folgenden der Verteilung

$$f : \mathbb{R} \to \mathbb{R}, \quad f(y) := \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad \sigma \in \mathbb{R}^+, \mu \in \mathbb{R}\,. \tag{2.1.1}$$

Hierbei ist allerdings noch nicht berücksichtigt, welcher Genotyp dem Phänotyp zugrundeliegt. Wir machen folgende Annahmen:

1. Es handelt sich um einen "single-locus trait", der genetische Einfluss auf den quantitativen Phänotyp ist also im Wesentlichen durch einen Genort bestimmt.

2. Wir können die Allele am Genort zu den Klassen "+" und "m" zusammenfassen, wobei "m" für die den Phänotyp verändernden Mutationsallele steht und "+" für die restlichen Allele, die dem Wildtyp entsprechen.

Damit ergeben sich für eine Person die Genotypkombinationen $(+,+)$, $(m,+) = (+,m)$ und (m,m). Sollte Imprinting vorliegen, wird noch zwischen $(m,+)$ und $(+,m)$ unterschieden. Das paternal geerbte Allel wird dabei zuerst aufgeführt. Wir nehmen nun an, dass der Phänotyp für jede dieser Genotypkombinationen normalverteilt ist, und zwar jede mit eigener Standardabweichung σ_i und eigenem Erwartungswert μ_i. Das heißt, der Genotyp bestimmt eine "Tendenz" des Phänotyps (= Erwartungswert μ_i), die restlichen Schwankungen sind auf Umwelteinflüsse und eventuelle weitere, in unserem Modell nicht explizit berücksichtigte Gene zurückzuführen. (2.1.1) erweitert sich wie folgt:

Definition 2.1.1. Es sei G die Menge aller Genotypkombinationen, d. h.

$$G := \{(+,+), (m,+) = (+,m), (m,m)\} \tag{2.1.2}$$

bzw.

$$G := \{(+,+), (m,+), (+,m), (m,m)\} \tag{2.1.3}$$

bei Imprinting. Weiter seien

$$f_i : \mathbb{R} \to \mathbb{R}, \quad f_i(y) := \frac{1}{(2\pi\sigma_i^2)^{\frac{1}{2}}} \exp\left(-\frac{(y-\mu_i)^2}{2\sigma_i^2}\right), \quad \sigma_i \in \mathbb{R}^+, \mu_i \in \mathbb{R}\,, \tag{2.1.4}$$

die zu den Genotypkombinationen gehörenden Normalverteilungen, mit $i \in G$.

Da die Unterscheidung zwischen $(m,+)$ und $(+,m)$ ohne Modellierung von Imprinting nicht vonnöten ist, bezeichnen wir diesen Genotyp einfach ohne Einschränkung mit $(m,+)$ oder mit (het) (heterozygot).

Das Mutationsallel kann die Tendenz des Phänotyps entweder erhöhen oder absenken. In einem vollständig dominanten oder rezessiven Erbgang fallen zwei der Dichtefunktionen zu einer zusammen: Entweder $f_{(m,+)}$ und $f_{(m,m)}$ oder $f_{(m,+)}$ und $f_{(+,+)}$ sind dann identisch. Bild 2.1 soll das Konzept der Dichtefunktionen noch einmal veranschaulichen.

Wie schon erwähnt, lassen sich Wahrscheinlichkeiten für einen Phänotyp y nicht angeben. Mit Def. 2.1.1 lässt sich allerdings bestimmen, wie wahrscheinlich es ist, dass

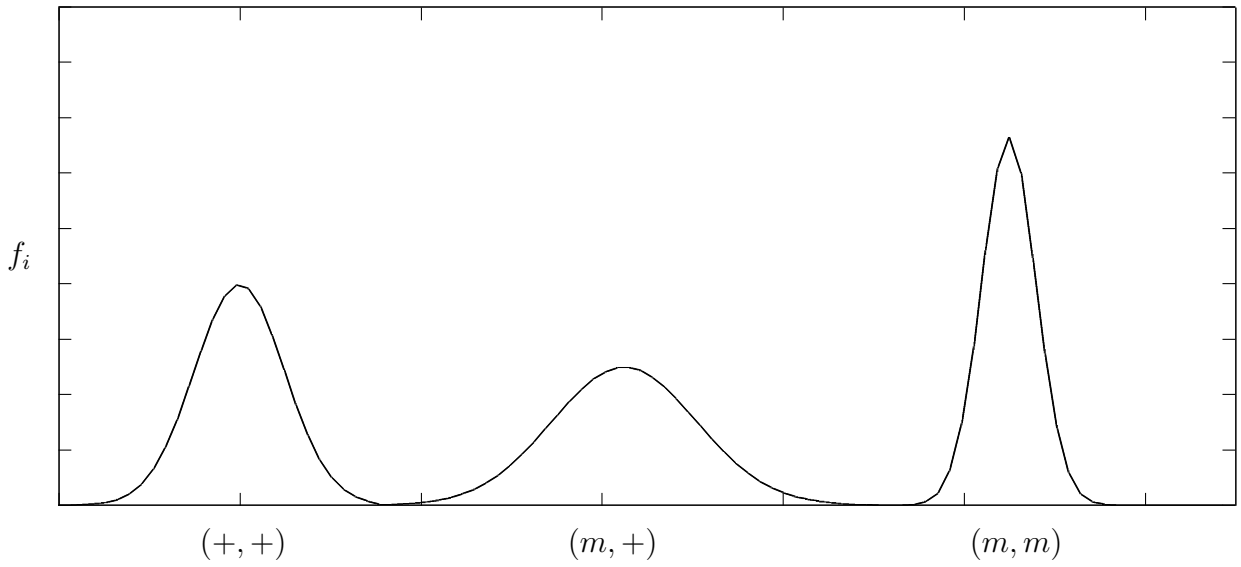

Bild 2.1: Dichtefunktionen des Phänotyps

eine Person, bedingt auf einen speziellen Genotyp i, einen quantitativen Phänotyp y in einem gewissen *Intervall* $[y_l, y_u]$ besitzt. Es handelt sich hierbei einfach um das Integral über die entsprechende Dichtefunktion:

$$P\left(y \in [y_l, y_u] \,|\, i\,\right) = \int_{y_l}^{y_u} f_i(y)\,\mathrm{d}y = \int_{y_l}^{y_u} \frac{1}{(2\pi\sigma_i^2)^{\frac{1}{2}}} \exp\left(-\frac{(y-\mu_i)^2}{2\sigma_i^2}\right)\,\mathrm{d}y. \qquad (2.1.5)$$

(2.1.5) wird im Folgenden die "klassischen" Penetranzen der dichotomen Kopplungsanalyse ersetzen.

2.2 Kopplungsanalyse mit quantitativen Phänotypen

In diesem Unterkapitel wird kurz erläutert, wie sich die Berechnung des LOD-Scores ändert, wenn man quantitative anstatt dichotomer Phänotypen analysieren möchte. Wie bereits erwähnt, beschränken wir uns auf die LOD-Score-Berechnung, die in GENEHUNTER verwendet wird. Hier gilt nach Strauch (2002), S. 53:

$$S(v, \phi^d) := LR(v) = \frac{\sum_{g^d} P\left(g^d \mid v\right) \prod_{i=1}^{k} \text{Pen}\left(\phi_i^d \mid g_i^d\right)}{\sum_{w \in \mathcal{V}} \sum_{g^d} P\left(g^d \mid w\right) \prod_{i=1}^{k} \text{Pen}\left(\phi_i^d \mid g_i^d\right) P_{\text{apriori}}(w)} \tag{2.2.1}$$

$\mathcal{V}$ bezeichnet die Menge aller möglichen Vererbungsvektoren für eine beliebige Stelle im Genom. Ein Vererbungsvektor ist ein binärer Vektor mit der Dimension der Zahl der effektiven Meiosen im Stammbaum. "0" bedeutet hierbei, dass eine Person das paternal geerbte Allel, "1", dass sie das maternal geerbte Allel weitervererbt hat. Es ist jeweils $v, w \in \mathcal{V}$. "k" bezeichnet die Anzahl der Personen im untersuchten Stammbaum. Der Vektor $\phi^d = \left(\phi_1^d, \ldots, \phi_k^d\right)$ fasst die Krankheitsphänotypen aller Individuen zusammen. Weiterhin ist $g^d = \left(g_1^d, \ldots, g_k^d\right)$ eine bestimmte Kombination von Genotypen am Krankheitslocus für alle Personen im Stammbaum. $\text{Pen}\left(\phi_i^d \mid g_i^d\right)$ ist die Wahrscheinlichkeit für die i-te Person, den Phänotyp ϕ_i^d zu besitzen, falls der Genotyp g_i^d vorliegt, vgl. Penetranzen aus Kapitel 1.2. $P\left(g^d \mid v\right)$ schließlich gibt die Wahrscheinlichkeit für eine Genotypkombination bedingt auf einen Vererbungsvektor v an. Durch diesen Term werden mittels der Einträge ungleich null alle Genotypkombinationen herausgegriffen, die mit v kompatibel sind.

Der LOD-Score berechnet sich dann mittels Summation über alle $w \in \mathcal{V}$ und Gewichtung mit $P(w)$:

$$\overline{S}\left(x, \phi^d\right) := \sum_{w \in \mathcal{V}} S\left(w, \phi^d\right) P\left(V(x) = w\right). \tag{2.2.2}$$

$P\left(V(x) = w\right)$ gibt dabei die Wahrscheinlichkeit an, dass an Position x der entsprechende Marker mittels w vererbt wurde. Es ist dann

$$\text{LOD} = \log_{10} \overline{S}\left(x, \phi^d\right). \tag{2.2.3}$$

Im quantitativen Fall müssen nun die Penetranzen $\text{Pen}\left(\phi_i^d \mid g_i^d\right)$ in (2.2.1) durch die Wahrscheinlichkeiten in (2.1.5) ersetzt werden. Hier stellt sich allerdings die Frage nach den Intervallgrenzen y_l und y_u. Da man im Prinzip die Wahrscheinlichkeit für die beobachteten Daten, also neben den Markergenotypen den Phänotyp y, bestimmen möchte, wäre eine erste Idee, das Intervall um den beobachteten Phänotyp der entsprechenden Person möglichst klein zu wählen. Mit einem kleinen Trick kann man die Intervallgrenzen jedoch auch ganz verschwinden lassen:

Für zwei Phänotypen y_j und y_t betrachten wir die beiden Terme

$$\int_{y_{l_j}}^{y_{u_j}} f_i(y)\,\mathrm{d}y \quad \text{sowie} \int_{y_{l_t}}^{y_{u_t}} f_{i'}(y)\,\mathrm{d}y. \tag{2.2.4}$$

Die Dichtefunktionen müssen nicht notwendig gleich sein, deswegen die Unterscheidung zwischen i und i'.

Für die Intervallgrenzen gelte

$$y_{l_j} := y_j - \varepsilon,\ y_{u_j} := y_j + \varepsilon \tag{2.2.5}$$

bzw.

$$y_{l_t} := y_t - \varepsilon,\ y_{u_t} := y_t + \varepsilon. \tag{2.2.6}$$

Nach dem Mittelwertsatz gilt dann $\forall \varepsilon > 0$

$$\frac{\int_{y_{l_j}}^{y_{u_j}} f_i(y)\,\mathrm{d}y}{\int_{y_{l_t}}^{y_{u_t}} f_{i'}(y)\,\mathrm{d}y} = \frac{2\varepsilon f_i(\hat{y_j})}{2\varepsilon f_{i'}(\hat{y_t})} = \frac{f_i(\hat{y_j})}{f_{i'}(\hat{y_t})}, \tag{2.2.7}$$

mit $\hat{y_j} \in \left[y_{l_j}, y_{u_j}\right]$ und $\hat{y_t} \in [y_{l_t}, y_{u_t}]$. Offensichtlich gilt

$$\hat{y_j} \to y_j, \quad \hat{y_t} \to y_t \tag{2.2.8}$$

für $\varepsilon \to 0$. Da f_i stetig ist, gilt dann mit (2.2.7)

$$\lim_{\varepsilon \to 0} \frac{\int_{y_{l_j}}^{y_{u_j}} f_i(y)\,\mathrm{d}y}{\int_{y_{l_t}}^{y_{u_t}} f_{i'}(y)\,\mathrm{d}y} = \frac{\lim_{\varepsilon\to 0} f_i(\hat{y_j})}{\lim_{\varepsilon\to 0} f_{i'}(\hat{y_t})} = \frac{f_i(\lim_{\varepsilon\to 0} \hat{y_j})}{f_{i'}(\lim_{\varepsilon\to 0} \hat{y_t})} = \frac{f_i(y_j)}{f_{i'}(y_t)}. \tag{2.2.9}$$

Da die Penetranzen in (2.2.1), die ja durch die Wahrscheinlichkeiten in (2.1.5) ersetzt werden sollen, im Rahmen der LOD-Score-Berechnung ebenfalls ausschließlich als Quotienten auftreten, ist es sinnvoll, diese gleich durch die Dichtefunktionen f_i zu ersetzen. (2.2.1) wird also im quantitativen Fall zu

$$S(v, \phi^d) := LR(v) = \frac{\sum_{g^d} P\left(g^d \mid v\right) \prod_{i=1}^{k} f_{g_i^d}(y_i)}{\sum_{w\in\mathcal{V}} \sum_{g^d} P\left(g^d \mid w\right) \prod_{i=1}^{k} f_{g_i^d}(y_i) P_{\text{apriori}}(w)}. \tag{2.2.10}$$

3 MOD-Score-Analyse

So wie im dichotomen Fall die Kenntnis der Penetranzen und der Krankheitsallelfrequenz vonnöten ist, um eine Kopplungsanalyse durchzuführen, benötigt man bei quantitativen Phänotypen Informationen über die Erwartungswerte μ_i und die Standardabweichungen σ_i der Dichtefunktionen f_i, $i \in G$, siehe Def. 2.1.1. Diese sind jedoch in den meisten Fällen unbekannt. Man kann sich hier wie im dichotomen Fall mit einer Maximierung des LOD-Scores über alle Parameter des Krankheitsmodells behelfen, in diesem Falle μ_i, σ_i sowie die Krankheitsallelfrequenz p. Die Parameterkombination mit dem maximalen LOD-Score, der nun als MOD-Score bezeichnet wird, gilt dann als die plausibelste für das Krankheitsmodell. Den Prozess der LOD-Score-Maximierung über die Parameter des Krankheitsmodells bezeichnet man als MOD-Score-Analyse.

3.1 Grundidee

Mathematisch handelt es sich bei der LOD-Score-Maximierung um eine Extremwertbestimmung einer mehrdimensionalen reellwertigen Funktion

$$g : \mathbb{R}^n \to \mathbb{R}, \quad w \mapsto g(w), \tag{3.1.1}$$

wobei

$$\begin{aligned} w &= (w_1, w_2, w_3, w_4, w_5, w_6, w_7) \\ &= \big(\mu_{(+,+)}, \mu_{(m,+)}, \mu_{(m,m)}, \sigma_{(+,+)}, \sigma_{(m,+)}, \sigma_{(m,m)}, p\big) \end{aligned} \tag{3.1.2}$$

bzw.

$$\begin{aligned} w &= (w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9) \\ &= \big(\mu_{(+,+)}, \mu_{(m,+)}, \mu_{(+,m)}, \mu_{(m,m)}, \sigma_{(+,+)}, \sigma_{(m,+)}, \sigma_{(+,m)}, \sigma_{(m,m)}, p\big) \end{aligned} \tag{3.1.3}$$

den Variablenvektor darstellt. n ist also 7, im Imprintingfall 9. g bezeichnet den über θ maximierten LOD-Score, die Argumente der Funktion sind die Krankheitsmodellparameter. Meistens lassen sich für die gesuchten Variablen vor der eigentlichen Maximierung obere und untere Schranken finden. So ergeben sich oft aus einfacher Logik Grenzen (physiologische Parameter können nicht unter 0 sinken). Andere sinnvolle Werte lassen sich eventuell durch biologisches oder medizinisches Hintergrundwissen definieren (höchster je gemessener Cholesterinspiegel als obere Schranke o. ä.). Hat man für jede Variable eine obere und untere Schranke gefunden, d. h. existieren L_i, U_i mit

$$L_i \leq w_i \leq U_i, \quad i = 1, \ldots, n, \tag{3.1.4}$$

so reduziert sich die Suche im $\mathbb{R}^n$ auf eine Suche in einem n-dimensionalen Quader

$$Q := [L_1, U_1] \times \cdots \times [L_n, U_n]. \tag{3.1.5}$$

Wir bezeichnen mit

$$L := (L_1, \ldots, L_n) \text{ sowie } U := (U_1, \ldots, U_n) \tag{3.1.6}$$

die Vektoren der unteren und oberen Schranken. Wir suchen also

$$w_{\max} \in Q \quad \text{mit} \quad g(w_{\max}) = \max_{w \in Q} g(w). \tag{3.1.7}$$

Da Q kompakt und g stetig, sogar differenzierbar ist, wissen wir, dass sowohl Maximum als auch Minimum von g auf Q existieren, vgl. z. B. Heuser (2004) und Heuser (2006) oder ein beliebiges anderes Standardwerk der Analysis. Es gibt also ein $w_{\max}$, welches (3.1.7) erfüllt.

3.2 Methoden der Optimierung

Wie bestimmt man nun ein solches $w_{\max}$? Um Methoden zu beschreiben, die einen solchen Wert finden sollen, ist zunächst die Klärung einiger Begrifflichkeiten vonnöten:

Definition 3.2.1. Sei $g \in C^1(Q, \mathbb{R})$, d. h. g ist in jeder Variable stetig und einmal

differenzierbar. Wir bezeichnen mit

$$\partial_{w_i} g(w) := \frac{\partial g}{\partial w_i}(w) \tag{3.2.1}$$

die *i-te partielle Ableitung* von g in w. Weiterhin sei

$$\nabla g : Q \to \mathbb{R}^n, \quad \nabla g(w) = (\partial_{w_1} g(w), \ldots, \partial_{w_n} g(w)) \tag{3.2.2}$$

der *Gradient* von g in w. Schließlich sei

$$\|\nabla g(w)\|_2 := \sqrt{(\partial_{w_1} g(w))^2 + \cdots + (\partial_{w_n} g(w))^2} \tag{3.2.3}$$

die *euklidische Norm* des Gradienten von g in w.

Bemerkung 3.2.2. Der Vektor $\nabla g(w)$ zeigt stets in Richtung des steilsten Anstiegs von g in w. Stellt man sich $\nabla g(w)$ als Pfeil im Raum vor, so gibt $\|\nabla g(w)\|_2$ die Länge des Pfeils an. $\|\nabla g(w)\|_2$ ist genau dann 0, wenn $\nabla g(w)$ in jeder Komponente 0 ist, d. h. wenn jede partielle Ableitung $\partial_{w_i} g(w)$ Null ist. Dies ist äquivalent dazu, dass g in w in keine Richtung ansteigt oder abfällt.

Nach obiger Bemerkung sind Stellen mit $\|\nabla g(w)\|_2 = 0$ aussichtsreiche Kandidaten für Extremstellen. Hier liegt entweder ein lokales Maximum, Minimum oder ein Sattelpunkt vor. Um diese Stellen zu identifizieren, müssen wir, ebenfalls nach Bemerkung 3.2.2, zunächst die Nullstellen aller ersten partiellen Ableitungen finden. Unglücklicherweise handelt es sich dabei jedoch um sehr komplizierte Funktionen, so dass die Nullstellensuche nicht analytisch zu bewerkstelligen ist. Es müssen also zwangsläufig Näherungslösungen der Nullstellen berechnet werden, wofür wiederum numerische Verfahren bemüht werden. Erschwerend kommt hinzu, dass der Extremwert auf einer beschränkten Menge gesucht wird, das Maximum also auch auf dem Rand von Q liegen könnte. Dies ist in der Untersuchung der w mit $\nabla g(w) = 0$ noch nicht enthalten und muss gesondert betrachtet werden. Man kann sich also gleich bei der Suche des Maximums eines Verfahrens bedienen, das eine Näherungslösung von $w_{\max}$ berechnet. Eine Bedingung an das Verfahren ist, dass die Näherungslösung von $w_{\max}$ in akzeptabler Zeit berechnet wird. Dies bedeutet, dass sich der Rechenaufwand des Verfahrens in Grenzen halten muss. Insbesondere bei Datensätzen mit großen Familien ist die Berechnung des LOD-Scores in GENEHUNTER extrem aufwendig, was zu langen Rechenzeiten führt. Es gilt also, einen Kompromiss zwischen

der Genauigkeit des Ergebnisses und der Berechnungsgeschwindigkeit zu finden.

Numerische Verfahren zur Optimierung mehrdimensionaler Probleme stellen die moderne Mathematik und die Informatik in großer Zahl zur Verfügung. Es gibt eine Fülle von Verfahren, die sich je nach Problemstellung unterschiedlich gut eignen. Eine kleine, aber feine Auswahl an iterativen Verfahren, sowohl für beschränkte als auch unbeschränkte Probleme, ist z. B. in Kelley (1999) zu finden. In SIAM (1996) wird man ebenfalls fündig.

Bei den gradienten-freien Methoden wäre hier z. B. der *Downhill-Simplex*-Algorithmus (*Nelder-Mead*-Algorithmus) zu nennen. Auf das Problem der LOD-Score-Optimierung angewandt, konvergierte dieser jedoch äußerst selten. Auch das Resultat der Optimierung unterschied sich kaum von den benutzten Startwerten. Eine weitere Methode ist das *simulated annealing*, welches sich jedoch als zu rechenaufwendig herausstellte, um bei dieser Anwendung von praktischem Nutzen zu sein. Die Anzahl der Funktionsauswertungen, die der Algorithmus benötigte, um zu konvergieren, war deutlich zu hoch, als dass das Problem in akzeptabler Rechenzeit gelöst werden konnte. Unter den gradienten-basierten Verfahren testeten wir die "Spectral Projected Gradient Method" (Birgin *et al.*, 2000) und das "GENCAN" Verfahren (Birgin & Martínez, 2002). Beide Verfahren lieferten sowohl hinsichtlich Konvergenz als auch Rechenzeit akzeptable Resultate. Allerdings wurden die Verfahren in beiden Punkten von dem "projizierten Gradienten-Verfahren" (im folgenden PGRAD genannt) übertroffen, welches letztendlich für die LOD-Score-Maximierung verwendet wurde. Das Verfahren ist ebenfalls in Kelley (1999) zu finden.

Das PGRAD-Verfahren basiert auf dem sog. Steepest-Descent-Algorithm und ist von seinem Grundprinzip recht anschaulich. Es wird im Folgenden erläutert. Es sei erwähnt, dass sämtliche Verfahren darauf ausgelegt sind, das Minimum und nicht das Maximum einer Funktion h zu finden. Dies stellt jedoch für unsere Anwendung kein Hindernis dar, wir müssen lediglich das Minimum von $-h$ berechnen, welches dann dem Maximum von h entspricht. Grundgedanke des folgenden Verfahrens ist, dass man eine Folge von Werten $w^0, w^1, ... \in Q$ konstruiert, mit denen man immer weiter in das Minimum "hinein läuft". Den $i+1$-ten Wert der Folge konstruiert man dabei mit dem i-ten Wert (Iteration), wobei man versucht, von w^i aus ein Stück weit in Richtung des steilsten Abfalls der Funktion h in w^i zu gehen.

Um das Minimum einer Funktion h zu finden, muss man zunächst einen Startwert $w^0 \in Q$ wählen. Theoretisch kann man jeden beliebigen Punkt in Q als Startwert

nehmen, es empfiehlt sich jedoch, einige Vorüberlegungen anzustellen, um bereits möglichst nahe an das Minimum heranzukommen. Je näher der Startwert bereits bei dem tatsächlichen Minimum $w_{\min}$ liegt, desto schneller und effizienter konvergiert das Verfahren.

Um die $i+1$-te Iterierte w^{i+1} aus der i-ten Ierierten w^i zu erhalten, berechnet man zunächst den Gradienten an der Stelle w^i:

$$\nabla h(w^i) = \left(\frac{\partial h}{\partial w_1}(w^i), \ldots, \frac{\partial h}{\partial w_n}(w^i) \right). \tag{3.2.4}$$

Wie bereits erwähnt beschreibt der Gradient die Richtung des steilsten Anstiegs, $-\nabla h(w^i)$ entsprechend die Richtung des steilsten Abfalls von h in w^i. Als nächstes führen wir den Projektor $P : \mathbb{R}^n \to Q$ auf Q ein. Für $w \in \mathbb{R}^n$ gilt

$$P(w)_j = \begin{cases} L_j & w_j \leq L_j \\ w_j & L_j < w_j < U_j \\ U_j & w_j \geq U_j. \end{cases} \tag{3.2.5}$$

P wird benutzt, um sicherzustellen, dass die Iteration innerhalb der spezifizierten Grenzen bleibt. Nun wird die nächste Iterierte berechnet:

$$w^{i+1} := P\left(w^i - \lambda_i \nabla h\left(w^i\right)\right). \tag{3.2.6}$$

$\lambda_i > 0$ ist ein sogenannter *Schrittlängenparameter*. Da $-\nabla h\left(w^i\right)$ die Richtung des steilsten Abfalls angibt, können wir (3.2.6) als einen Schritt der Länge $\lambda_i ||\nabla h\left(w^i\right)||$ in Richtung des Minimums von h interpretieren. Die Wahl von λ_i geschieht nach folgendem Kriterium: Zunächst definieren wir

$$w(\lambda_i) = P\left(w - \lambda_i \nabla h\left(w\right)\right). \tag{3.2.7}$$

Dann ist λ_i ein gültiger Schrittlängenparameter, wenn

$$h\left(w(\lambda_i)\right) - h(w) \leq \frac{-\delta}{\lambda_i} ||w - w(\lambda_i)||^2 \tag{3.2.8}$$

gilt. $\delta > 0$ ist hierbei ein frei wählbarer Parameter des Algorithmus. Typischerweise wird er auf 10^{-4} gesetzt. Gleichung (3.2.8) nennt man auch "*condition of sufficient decrease*". Sie läßt sich wie folgt interpretieren: Nehmen wir an, dass das Argument von P in Gleichung (3.2.7) in Q liegt, und setzen (3.2.7) in (3.2.8) ein, so erhalten

wir

$$\begin{aligned} h\left(w(\lambda_i)\right) - h(w) &\leq \frac{-\delta}{\lambda_i} \left|\left|\lambda_i \nabla h\left(w\right)\right|\right|^2 \\ &= \frac{-\delta\lambda_i^2}{\lambda_i} \left|\left|\nabla h\left(w\right)\right|\right|^2 \\ &= -\delta\lambda_i \left|\left|\nabla h\left(w\right)\right|\right|^2. \end{aligned} \tag{3.2.9}$$

Das heißt, durch die Wahl von λ_i muss sich die Funktion h beim Iterationsschritt von w^i nach w^{i+1} mindestens um den Betrag $\delta\lambda_i \left|\left|\nabla h\left(w\right)\right|\right|^2$ verringern.

Die Iteration (3.2.6) wird nun so lange fortgesetzt, bis ein vorher definiertes Abbruchkriterium greift. Dies kann eine maximale Anzahl an Iterationen sein, oder ein Kriterium für ein lokales bzw. globales Minimum. Dies wird später noch ausführlicher erläutert. Will man das PGRAD-Verfahren auf das Problem der MOD-Score-Berechnung mit der LOD-Score-Funktion g anwenden, startet man den Algorithmus mit $h = -g$ als Argument. Bild 3.1 soll die Funktionsweise des Verfahrens noch einmal veranschaulichen.

Von entscheidendem Einfluss auf den Verlauf der Iteration ist hierbei der Startwert. Liegt dieser zu weit von dem gesuchten Minimum entfernt, so ist es gut möglich, dass die Iteration in eine falsche Richtung läuft und das Minimum "verpasst". Ebenfalls möglich ist, dass die Iteration in ein kleineres, lokales Minimum hinein läuft. Der Wahl des Startwertes kommt also eine besondere Bedeutung zu, weswegen ihr ein eigenes Unterkapitel gewidmet wird.

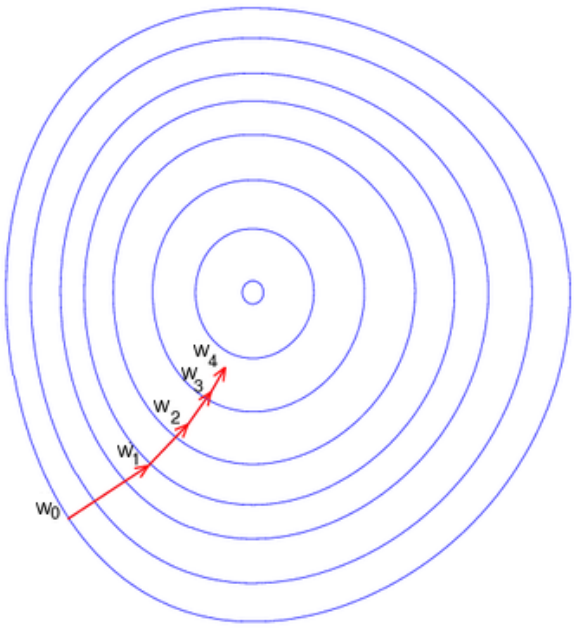

Bild 3.1: PGRAD-Verfahren
Die Linien entsprechen Höhenstufen der Funktion g, ähnlich einer topographischen Karte. Das Maximum befindet sich in der Mitte. Die Pfeile symbolisieren den Gradienten, der dazu benutzt wird, den nächsten Punkt der Iteration zu berechnen. (Quelle: Wikipedia)

3.3 Startwertbestimmung

Das PGRAD-Verfahren kann prinzipiell mit jedem Wert $w_0 \in Q$ als Startwert arbeiten. Allerdings führt nicht jeder Startwert zu einem gleich guten Ergebnis der Iteration. Je näher der gewählte Startwert bei dem tatsächlichen Maximum liegt, desto schneller und sicherer konvergiert das Verfahren. Der Grund dafür ist mit Bild 3.1 leicht ersichtlich: Liegt w_0 nahe am Optimum, sind nur wenige Schritte vonnöten, um dieses auch zu erreichen. Weiter entfernte Startwerte erhöhen die Anzahl der Iterationsschritte, was zu längeren Rechenzeiten führt. Außerdem können entfernte Startwerte dazu führen, dass die Iteration in ein lokales Nebenmaximum hinein läuft und dort beendet wird. Die Wahl des Startwertes beeinflusst also zum einen die Geschwindigkeit und zum anderen die Wahrscheinlichkeit, das gesuchte Maximum auch tatsächlich zu finden. Es empfiehlt sich daher, einige Vorüberlegungen bei der Wahl des Startwertes anzustellen. Die folgenden Methoden zur Startwertbestimmung haben sich als effektiv erwiesen, kommen deswegen in GENEHUNTER-QMOD zur An-

wendung und sind darum hier beschrieben. Sie sind jedoch keinesfalls die einzigen Möglichkeiten, da das PGRAD-Verfahren wie erwähnt mit jedem Startwert arbeiten kann.

Es sei k die Anzahl aller untersuchten Personen und $q \in \mathbb{R}^k$ der Vektor der Phänotypen der Personen, d.h. q_i ist der Phänotyp der i-ten Person im Stammbaum, $i = 1, \ldots, k$. Die Startwerte des PGRAD-Verfahrens berechnen sich bei den hier benutzten Methoden allesamt mit den Daten aus q. Es sei hier noch einmal erwähnt, dass es sich bei den Startwerten um (sehr grobe) Schätzer für $\mu_{(+,+)}, \mu_{(m,+)}, \mu_{(m,m)}, \sigma_{(+,+)}, \sigma_{(m,+)}, \sigma_{(m,m)}$ und die Krankheitsallelfrequenz p handelt. Letztere wird in den Startwertberechnungen standardmäßig auf 10^{-1} gesetzt. Nützlich ist noch die Überlegung, ob das Mutationsallel m eine Erhöhung oder ein Absinken des quantitativen Phänotyps zur Folge hat. Weiterhin empfiehlt es sich, den Datensatz nach der Größe zu ordnen, sei also ohne Einschränkung $q_i \leq q_j,\ i \leq j$.

Methode 1:

Wir teilen den Datensatz in das untere, mittlere und obere Drittel, d. h. bei den Indizes $[\frac{k}{3}]$ und $[\frac{2k}{3}]$, auf. [ ] bedeutet hier, dass der Bruch immer zur nächsten natürlichen Zahl gerundet wird. Wir berechnen separat den empirischen Mittelwert und die empirische Standardabweichung für jeden der drei Teile, und benutzen die so erhaltenen Werte als Startwerte.

$$\begin{aligned}
\mu_{(+,+)} &:= \frac{1}{[\frac{k}{3}]} \sum_{i=1}^{[\frac{k}{3}]} q_i, \qquad & \sigma_{(+,+)} &:= \frac{1}{[\frac{k}{3}]-1} \sum_{i=1}^{[\frac{k}{3}]} \left(q_i - \mu_{(+,+)}\right)^2 \\
\mu_{(m,+)} &:= \frac{1}{[\frac{k}{3}]} \sum_{i=[\frac{k}{3}]+1}^{[2\frac{k}{3}]} q_i, & \sigma_{(m,+)} &:= \frac{1}{[\frac{k}{3}]-1} \sum_{i=[\frac{k}{3}]+1}^{[2\frac{k}{3}]} \left(q_i - \mu_{(m,+)}\right)^2 \\
\mu_{(m,m)} &:= \frac{1}{[\frac{k}{3}]} \sum_{i=[2\frac{k}{3}]+1}^{k} q_i, & \sigma_{(m,m)} &:= \frac{1}{[\frac{k}{3}]-1} \sum_{i=[2\frac{k}{3}]+1}^{k} \left(q_i - \mu_{(m,m)}\right)^2
\end{aligned} \tag{3.3.1}$$

Hierbei wurde angenommen, dass Genotypen mit dem Mutationsallel einen höheren Phänotyp hervorrufen. Die Idee hinter dieser Startwertberechnung lässt sich durch Abbildung 2.1 veranschaulichen. Kleinere Phänotypen werden mit höherer Wahrscheinlichkeit beobachtet, wenn man den Genotyp $(+,+)$ zugrunde legt, entsprechend mittelgroße und große Phänotypen bei Annahme von $(m,+)$ und (m,m). Wir

nehmen deswegen an, das erste Drittel aller Werte sei eine Stichprobe von $f_{(+,+)}$, das zweite von $f_{(m,+)}$ und das letzte Drittel von $f_{(m,m)}$. Aus den Stichproben konstruieren wir dann mittels (3.3.1) grobe Schätzer für die Parameter der Normalverteilungen, die zu optimieren sind. Erwartet man eine Verringerung des Phänotyps durch das Mutationsallel, muss man die Zuweisung von $\mu_{(+,+)}$ und $\mu_{(m,m)}$ vertauschen. Im Falle von Imprinting wird $\mu_{(m,+)} = \mu_{(+,m)}$ gesetzt.

Methode 2:

Die Berechnung ist die gleiche wie in Methode 1, nur dass wir den Datensatz an anderen Stellen aufteilen. Wir suchen uns die beiden größten Differenzen aufeinander folgender Phänotypen im Datensatz und schneiden dort ab. Wir suchen also j_1 mit

$$q_{j_1+1} - q_{j_1} = \max_{j=1,\ldots,k-1} (q_{j+1} - q_j) \tag{3.3.2}$$

sowie j_2 mit

$$q_{j_2+1} - q_{j_2} = \max_{j=1,\ldots,k-1, j\neq j_1} (q_{j+1} - q_j). \tag{3.3.3}$$

Die Berechnung der Startwerte erfolgt dann analog zu (3.3.1), nur dass bis zu j_1 und j_2 summiert wird und nicht bis $[\frac{k}{3}]$ und $[\frac{2k}{3}]$.

Methode 3: Eine weitere Möglichkeit zur Startwertbestimmung erhält man über die Quartile des Datensatzes. Wir berechnen $Q_{0.25}$, $Q_{0.5}$ und $Q_{0.75}$ und setzen

$$\begin{aligned} \mu_{(+,+)} &:= Q_{0.25}, \\ \mu_{(m,+)} &:= Q_{0.5}, \\ \mu_{(m,m)} &:= Q_{0.75}. \end{aligned} \tag{3.3.4}$$

Sei weiterhin $\overline{q} := \frac{1}{k}\sum_{i=1}^{k} q_i$ der empirische Mittelwert des Datensatzes. Die Werte für $\sigma_{(+,+)}$, $\sigma_{(m,+)}$ und $\sigma_{(m,m)}$ erhalten wir durch Berechnung der empirischen Standardabweichung

$$s := \sqrt{\frac{1}{k-1}\sum_{i=1}^{k} (q_i - \overline{q})^2} \tag{3.3.5}$$

und setzen $\sigma_{(+,+)} := \sigma_{(m,+)} := \sigma_{(m,m)} := \frac{s}{3}$. Verringert die Mutation den Phänotyp, vertauscht man entsprechend $\mu_{(+,+)}$ mit $\mu_{(m,m)}$. Bei Imprinting setzt man $\mu_{(m,+)} = \mu_{(+,m)}$ sowie $\sigma_{(m,+)} = \sigma_{(+,m)}$.

Die Erfahrung zeigt, dass nicht jeder Datensatz mit jedem Startwertverfahren gleich gut funktioniert. Die hier vorgestellten Verfahren basieren auf der Annahme, dass der Genort einen hohen Anteil der Varianz erklärt. Bei Phänotypen mit großer umweltbedingter Streuung kann ein so gewählter Startwert u.U. eine ungünstige Ausgangsposition für eine Maximierung sein. Manchmal bricht die Iteration des PGRAD-Verfahrens schon nach wenigen Schritten mit einem MOD-Score von 0 ab. Empfehlenswert ist hier, die Iteration mit neuen Startwerten zu wiederholen. Sollte das Verfahren erneut früh abbrechen, muss ein dritter Startwert gewählt werden. Da es durchaus vorkommen kann, dass die Maximierung bei einem besonders "unglücklichen" Datensatz auch bei drei Startwerten nicht anschlägt, empfiehlt es sich, ein zusätzliches Verfahren zu konstruieren, das theoretisch unendlich viele Startwertkombinationen erzeugen kann, die solange auf die Iteration angesetzt werden, bis einer der Startwerte "greift". In der Praxis wird man im Hinblick auf Rechenzeit ein Limit an die maximale Anzahl von Startwerten setzen, die erzeugt werden. Aufgrund obiger Überlegungen wird nun eine Methode vorgestellt, die mit Hilfe eines Zufallsgenerators Startwerte erstellt und sich deswegen gut für die beschriebene Problematik eignet.

Methode 4:

Die Ausgangswerte für $\mu_{(+,+)}$, $\mu_{(m,+)}$ und $\mu_{(m,m)}$ seien wie in Methode 3 die Quartile des Datensatzes. Diese werden jetzt allerdings modifiziert. Dazu berechnen wir zuerst den Abstand des Quartils zu beiden Rändern des Datensatzes und nehmen den Kleineren von beiden. Diesen Wert gewichten wir mit einer Zufallsvariable zwischen 1 und -1 und addieren ihn zum Quartil. Formal: Seien $r_{(+,+)}$, $r_{(m,+)}$ und $r_{(m,m)}$ drei gleichverteilte Zufallszahlen aus $[-1, 1]$. Wir setzen als Startwerte für die Erwartungswerte

$$\begin{aligned} \mu_{(+,+)} &:= Q_{0.25} + r_{(+,+)} \min\left(Q_{0.25} - q_1, q_k - Q_{0.25}\right), \\ \mu_{(m,+)} &:= Q_{0.5} + r_{(m,+)} \min\left(Q_{0.5} - q_1, q_k - Q_{0.5}\right), \\ \mu_{(m,m)} &:= Q_{0.75} + r_{(m,m)} \min\left(Q_{0.75} - q_1, q_k - Q_{0.75}\right). \end{aligned} \tag{3.3.6}$$

Durch Addieren des gewichteten Minimalabstandes wird sichergestellt, dass die Erwartungswerte nicht über die Spanne des Datensatzes hinausgehen. Die Gleichverteilung der Zufallsvariablen sorgt dafür, dass sich der Startwert im Mittel auf dem jeweiligen Quartil befindet.

Die Standardabweichungen berechnen wir zunächst ebenfalls wie in Methode 3 mit Gleichung (3.3.5). Anschließend werden sie durch Multiplikation mit einer Zu-

fallsvariablen aus $(0, 2]$ gestreckt oder gestaucht. Seien $r_{(+,+)}$, $r_{(m,+)}$ und $r_{(m,m)}$ drei gleichverteilte Zufallszahlen aus $(0, 2]$. Dann setzen wir

$$\begin{aligned} \sigma_{(+,+)} &:= r_{(+,+)}\frac{s}{3}, \\ \sigma_{(m,+)} &:= r_{(m,+)}\frac{s}{3}, \\ \sigma_{(m,m)} &:= r_{(m,m)}\frac{s}{3}. \end{aligned} \tag{3.3.7}$$

Mit diesem Verfahren lassen sich theoretisch unendlich viele Startwertkombinationen erzeugen. Kombiniert man diese Methode mit dem PGRAD-Verfahren, kann man verschiedene Startwerte verwenden, bis die Maximierung einen MOD-Score größer null erreicht. In der Praxis legt man sich auf eine Maximalanzahl fest, die getestet wird. In der Implementierung wurden 10 Startwerte benutzt.

Ebenfalls möglich ist, den MOD-Score für eine feste Anzahl an Startwerten zu berechnen (z. B. zehn) und dann das Maximum dieser MOD-Scores als "endgültigen" MOD-Score zu benutzen. Auf diese Weise werden zum einen schlecht gewählte Startwerte ausgeglichen, deren Iterationen nach wenigen Schritten abbrechen. Zusätzlich wird die Wahrscheinlichkeit verringert, ein kleineres unbedeutendes Nebenmaximum als Ergebnis der Maximierung zu erhalten. Allerdings ist diese Variante aufwendiger, als Startwerte solange zu generieren, bis einer "greift". In der Implementierung wurde diese Variante benutzt. Es wurde je ein Startwert mit Methode 1,2 und 3 erzeugt, die restlichen sieben Startwerte wurden mit Methode 4 erzeugt.

3.4 Weitere Krankheitsmodelle

Das in Kapitel 2.1 vorgestellte Modell des Phänotyps ist sehr allgemein gehalten. Sämtliche Erwartungswerte μ_i sowie die Varianzen σ_i^2 können verschieden sein. Dies ist ein guter Ansatz, wenn über das Krankheitsmodell nicht viel bekannt ist bzw. die einzige Annahme darin besteht, dass der Phänotyp bedingt auf den Genotyp normalverteilt ist.

Es kommt jedoch vor, dass bereits vor der Analyse zusätzliche Informationen über das Krankheitsmodell vorliegen und ein derart allgemeiner Ansatz nicht vonnöten ist. So weiß man z.B. von vielen Phänotypen, dass im Erbgang keine Dominanz-Effekte auftreten. Dies bedeutet, dass der Erwartungswert $\mu_{(m,+)}$ der heterozygoten Personen dem arithmetischen Mittel der homozygoten Erwartungswerte entspricht:

$$\mu_{(m,+)} = \frac{1}{2}\left(\mu_{(+,+)} + \mu_{(m,m)}\right). \tag{3.4.1}$$

Der heterozygote Erwartungswert $\mu_{(m,+)}$ liegt also genau zwischen den homozygoten, d.h. sowohl das Wildtyp-Allel + als auch das Mutationsallel m besitzen das gleiche "Gewicht" bei der phänotypischen Ausprägung des heterozygoten Genotyps.

Weiß man um diese Information, kann und sollte man sie bei der Kopplungsanalyse berücksichtigen. Man passt das Modell aus Kap. 2.1 gemäß Gleichung (3.4.1) an. Da $\mu_{(m,+)}$ sich nun aus $\mu_{(+,+)}$ und $\mu_{(m,m)}$ ergibt, hat man einen Parameter weniger im Krankheitsmodell, der während der MOD-Score-Analyse auch nicht mehr geschätzt werden muss. Das PGRAD-Verfahren ist hierfür so anzupassen, dass bei der LOD-Score-Funktion der $\mu_{(m,+)}$-Parameter intern auf $\frac{1}{2}\left(\mu_{(+,+)} + \mu_{(m,m)}\right)$ gesetzt wird. Diese neue LOD-Score-Funktion besitzt dann einen Parameter weniger.

Eine ebenfalls nützliche Vereinfachung ist die der gleichen Varianzen. Oftmals kann angenommen werden, dass sowohl Umwelteinflüsse als auch weitere genetische Effekte den Phänotyp unabhängig vom Genotyp am Krankheitslocus gleichermaßen beeinflussen. Dies bedeutet, dass die Streuung (Varianz) um den mittleren Phänotyp μ_i für alle Genotypen $i \in \{(+,+),(m,+),(m,m)\}$ gleich ist. Wir machen also die Annahme

$$\sigma^2_{(+,+)} = \sigma^2_{(m,+)} = \sigma^2_{(m,m)}. \tag{3.4.2}$$

Hier spart man sich gegenüber dem allgemeinen Modell in 2.1 sogar zwei Parameter. Die Implementierung erfolgt analog zu (3.4.1).

Der Vollständigkeit halber sei noch einmal der Imprinting-Fall erwähnt. Hier wird bei heterozygoten Genotypen unterschieden, ob ein Allel paternal oder maternal geerbt worden ist. Es ergeben sich also im heterozygoten Fall die Genotypen $(m,+)$ und $(+,m)$, wobei das paternal geerbte Allel zuerst aufgeführt wird. Will man diesen Fall modellieren, benötigt man zwei statt einer Dichtefunktion für den heterozygoten Genotyp. Daraus ergeben sich zwei Parameter $\mu_{(m,+)}$ und $\mu_{(+,m)}$ statt $\mu_{(m,+)}$, sowie $\sigma_{(m,+)}$ und $\sigma_{(+,m)}$ statt $\sigma_{(m,+)}$, die in der MOD-Score-Analyse geschätzt werden müssen. Im Gegensatz zu (3.4.1) wird allerdings die Optimierung aufwendiger, da mehr Parameter geschätzt werden müssen.

Die vorgenannten Modellmodifikationen keiner Dominanzeffekte, gleicher Varianzen und Imprinting können selbstverständlich auch miteinander kombiniert werden.

Die Eingabeparameter der LOD-Score-Funktion müssen entsprechend modifiziert werden, danach übergibt man sie dem PGRAD-Verfahren.

Der Vorteil der verschiedenen Modelle ist, dass ein passendes Modell meist höhere MOD-Scores liefert. Eine ganze Klasse "falscher" Parametersätze, die durch die MOD-Score-Analyse theoretisch geschätzt werden könnten, werden von vornherein ausgeschlossen. Weiterhin reduziert sich in den Fällen keiner Dominanz-Effekte und gleicher Varianzen die Anzahl der zu schätzenden Parameter, was die Dimension des Optimierungsproblems verringert. Dies hat eine Verringerung der Rechenzeit zur Folge, die MOD-Score-Analyse wird dadurch schneller durchführbar. Durch die Reduzierung der Zahl der Freiheitsgrade, die damit einhergeht, können weiterhin genetische Effekte derselben Stärke mit größerer Wahrscheinlichkeit (Power) detektiert werden als mit einem Modell, welches unnötigerweise eine größere Zahl von Parametern enthält.

3.5 Abbruchkriterien

Wie bereits in Kap. 3.2 erwähnt, gilt bei lokalen Extrema $\|\nabla g\|_2 = 0$. Das Verschwinden des Gradienten liefert also ein sinnvolles Abbruchkriterium. Da allerdings auch moderne Computer nur mit endlicher Genauigkeit rechnen können, kann es durchaus vorkommen, dass der Gradient an einem Extremum geringfügig von 0 abweicht. Die Tatsache, dass innerhalb des Algorithmus gewisse Terme ebenfalls nur näherungsweise berechnet werden (z. B. die partiellen Ableitungen mittels finiter Differenzen), kann zu dem gleichen Effekt führen. Es empfiehlt sich also, die Iteration auch dann abzubrechen, wenn der Gradient nicht exakt 0 ist, sondern eine gewisse Toleranz unterschreitet. In der Anwendung mit GENEHUNTER-QMOD wurde diese Toleranz auf 10^{-3} gesetzt. Ein weiterer Punkt, der berücksichtigt werden muss, ist die maximale Anzahl an Iterationen, die das PGRAD-Verfahren durchführen darf. Im Hinblick auf Rechenzeit sollte auch hier eine obere Schranke gewählt werden. Eine Verringerung der Anzahl der erlaubten Iterationen verschlechtert immer die Qualität des Ergebnisses. Es zeigt sich allerdings, dass das PGRAD-Verfahren, falls der Startwert gut gewählt ist, ein (zumindest lokales) Maximum nach einigen hundert Iterationen zuverlässig findet und, falls es nicht von selbst abbricht, ein weiteres Iterieren den MOD-Score nur geringfügig weiter erhöht und die geschätzten Parameter sich kaum mehr ändern. Deshalb ist die Begrenzung der Iterationen auf ca. 500 keine starke

Einschränkung.

3.6 Struktur des Algorithmus

Im Folgenden wird mittels Pseudocode die Implementierung des eben besprochenen Optimierungsproblems skizziert. Im Wesentlichen sind drei Funktionen vonnöten. Die erste berechnet aus den Parametern des Krankheitsmodells, den Markergenotypen, den Markerallelfrequenzen, den Rekombinationsfrequenzen, den Verwandtschaftsbeziehungen und den quantitativen Phänotypen den (über alle genetische Positionen des Krankheitslocus maximierten) LOD-Score. Es handelt sich dabei um die Funktion g, vgl. (3.1.1).

function CALCULATE LOD-SCORE(parameter vector w, INPUT DATA)
  % berechne Maximum-LOD-Score nach Kap. 2.2.
  return maxlod;
end function

Die zweite Funktion berechnet aus der Menge aller quantitativen Phänotypen die Startwerte für das PGRAD-Verfahren. Hierbei kann z. B. eine integer-Variable übergeben werden, die bestimmt, nach welcher Methode die Startwerte erstellt werden:

function CALCULATE START VALUES(**int** i, INPUT DATA)
  % berechne Startwerte w mittels Methode i nach Kap. 3.3.
  % für $i > 3$ benutze Methode 4
  return w;
end function

Die letzte Funktion schließlich ist das PGRAD-Verfahren selbst. Als Argumente benötigt sie eine Startwertkombination, die zu maximierende Funktion g sowie die Randwerte des Parameterraums. Sollten sich diese in Abhängigkeit der Startwertberechnung und der Eingangsdaten nicht ändern, können sie auch im PGRAD-Verfahren fest gesetzt werden.

function PGRAD-OPTIMIZATION(start values w, function g, boundaries L, U)
% berechne Maximum von g auf Q wie in Kap. 3.2.
return maximum;
end function

Das eigentliche Programm hat dann folgende Struktur:

Algorithmus 3.1 MOD-Score-Berechnung mittels PGRAD

read INPUT DATA;
double maxlod = 0;
double save_maxlod = 0;
double[ ] w;

for $i = 1, \ldots, 10$ **do**
w = CALCULATE START VALUES(i, INPUT DATA);
maxlod = PGRAD-OPTIMIZATION(w, CALCULATE LOD-SCORE, L, U);
if (maxlod > save_maxlod) **then** save_maxlod = maxlod;
end if
end for

print save_maxlod;

4 p-Wert-Bestimmung und Testeigenschaften

Wie bereits in Kapitel 1.2 erwähnt, wird bei der Kopplungsanalyse ein statistischer Signifikanztest durchgeführt. Die Teststatistik ist im vorliegenden Fall der MOD-Score. Nachdem dieser berechnet worden ist, stellt sich die Frage, ob die Nullhypothese keiner Kopplung (H_0) verworfen werden darf. Dazu ist die Kenntnis des p-Wertes vonnöten. Bei dem p-Wert handelt es sich um die Wahrscheinlichkeit, unter der Annahme der Nullhypothese keiner Kopplung den aus den Daten berechneten MOD-Score zu erhalten, oder noch ein extremeres Ergebnis, also größeren MOD-Score, zu beobachten. Formal: Bezeichne z den MOD-Score und sei z_{R_0} der aus den Daten errechnete MOD-Score, dann ist

$$p := P\left(z \geq z_{R_0} \,|\, H_0\right) \tag{4.0.1}$$

der zugehörige p-Wert. Der MOD-Score gehorcht unter der Nullhypothese einer speziellen Verteilung, mittels der sich der p-Wert berechnen lässt. Leider ist diese Verteilung im Falle des PGRAD-Verfahrens unbekannt und der p-Wert damit nicht ohne Weiteres berechenbar. Die Verteilung des MOD-Scores kann aber durch Simulationen bestimmt werden.

4.1 Empirischer p-Wert

Um die Verteilung des MOD-Scores unter der Nullhypothese zu simulieren, ist es zunächst notwendig, aus den ursprünglichen Daten zufällig entsprechende Replikate unter der Nullhypothese keiner Kopplung zu konstruieren. Zu diesen sog. Nullhypothesenreplikaten berechnet man dann jeweils den MOD-Score. Bei genügend hoher Anzahl an Nullhypothesenreplikaten, die zusammen als Stichprobe aller möglicher Replikate unter H_0 zu verstehen sind, kann man so die Verteilungsfunktion des MOD-

Scores unter der Nullhypothese nachbilden.

Die Replikate unter H_0 werden dabei wie folgt erstellt: Die Familienstruktur, die Anzahl der Marker und der quantitative Phänotyp des Originaldatensatzes werden nicht verändert. Die Genotypen der Founder jedoch werden für jeden Marker per Zufall neu gesetzt, gemäß den Allelfrequenzen der entsprechenden Marker. Dann wird die Vererbung der Markerallele der Founder an die Nicht-Founder simuliert: Bei jeder Meiose an jedem Markerlocus wird jedes der beiden Allele mit 50%-iger Wahrscheinlichkeit weitervererbt. Die Vererbungswahrscheinlichkeit eines Allels an einem Markerlocus hängt also nicht von benachbarten Loci und deren genetischen Abständen ab. Jedes Allel fällt also ähnlich einer Kugel im Galton-Brett durch den Stammbaum. Man spricht auch von *gene dropping*. Hierbei wird jedwede Kopplung zwischen Phänotyp und Markerallelen zerstört, und man hat einen Datensatz unter der Nullhypothese keiner Kopplung generiert. Dies ist allerdings nur eine Möglichkeit der Simulation der Verteilung. Eine andere Methode wird z.B. in Zhao *et al.* (1999) beschrieben.

Mit dieser simulierten Verteilung lässt sich nun ein Schätzer für den p-Wert berechnen. Die Wahrscheinlichkeit, unter der Nullhypothese einen MOD-Score größer gleich dem beobachteten zu erhalten, entspricht ja der Wahrscheinlichkeit, dass ein MOD-Score eines Nullhypothesenreplikates den Original-MOD-Score übertrifft. Diese Wahrscheinlichkeit lässt sich aber gerade durch den relativen Anteil aller Nullhypothesenreplikate schätzen, deren MOD-Score über dem Original-MOD-Score liegt. Je größer die Anzahl der Replikate ist, desto genauer ist der Schätzer.

Wir führen nun den *empirischen p-Wert* ein:

Definition 4.1.1. Sei R_0 der Originaldatensatz mit MOD-Score z_{R_0}. Sei $\{R_1, \ldots, R_N\}$ die Menge der erstellten Nullhypothesenreplikate mit den MOD-Scores $z_{R_1}, \ldots, z_{R_N}$. Es sei weiter

$$N_0 := \#\{R_i \mid z_{R_i} \geq z_{R_0}\} \tag{4.1.1}$$

die Anzahl der Nullhypothesenreplikate, deren MOD-Score den der Originaldaten übersteigt. Dann ist

$$p := \frac{N_0}{N} \tag{4.1.2}$$

der *empirische p-Wert* des MOD-Scores z_{R_0} von R_0.

Der empirische p-Wert kann nun für die Testentscheidung benutzt werden. Liegt er unter dem vorher vorgegebenen Signifikanzniveau α von beispielsweise 5% (1% oder 0,1% ...), so gilt das Ergebnis als signifikant, und die Nullhypothese kann zugunsten der Alternativhypothese Kopplung (H_1) verworfen werden. Dies bedeutet, dass der Krankheitslocus des untersuchten Phänotyps an einen der Marker gekoppelt ist. Die Position des Locus ist durch die Rekombinationsfrequenz θ, im Multimarkerfall durch die genetische Position x gegeben, bei der sich der maximale MOD-Score ergibt. Die Testentscheidung kann wie folgt interpretiert werden: Übersteigt nur ein kleiner Teil (weniger als 5%, 1%, 0,1% ...) der MOD-Scores der Nullhypothesenreplikate den Original-MOD-Score, so ist es entsprechend unwahrscheinlich (gerade die Wahrscheinlichkeit p), dass ein solcher Score durch Zufall entstanden ist, falls die Nullhypothese gilt. Deswegen entscheidet man sich gegen die Nullhypothese keiner Kopplung. Das Signifikanzniveau α ist die Wahrscheinlichkeit für einen Fehler 1. Art, bei dem H_0 verworfen wird, obwohl H_0 gilt.

4.2 p-Wert-Berechnung mittels Funktionswert-Stichproben

Wie bereits erwähnt, handelt es sich bei dem empirischen p-Wert um einen *Schätzer* für die Wahrscheinlichkeit, unter H_0 den MOD-Score des Originaldatensatzes zu erreichen oder zu übertreffen. Er berechnet sich aus einer Stichprobe aus allen möglichen Nullhypothesenreplikaten. Je größer diese Stichprobe, desto genauer ist der Schätzer. Bei einer präzisen Berechnung des p-Wertes, insbesondere bei kleinen p-Werten, ist also eine große Anzahl von Replikaten im fünf- oder sechsstelligen Bereich erforderlich. Für jedes dieser Replikate muss der MOD-Score berechnet und das PGRAD-Verfahren mit unter Umständen mehreren Startwerten benutzt werden. Hier stellt sich die Frage nach der dafür benötigten Rechenzeit. Es hat sich gezeigt, dass bei einer gründlichen Maximierung (ca. zehn Startwerte) die p-Wert-Berechnung schon bei kleinen Familienstammbäumen erheblich ist (zwei Tage für 5000 Replikate, 1 Prozessor auf dem Marburger Rechencluster (MaRC)). Die Frage nach einer Alternative ist gerechtfertigt.

Glücklicherweise benötigen wir bei der p-Wert-Berechnung den MOD-Score streng genommen gar nicht. Es ist lediglich von Interesse, ob der MOD-Score des realen Datensatzes größer ist als der des Nullhypothesenreplikates oder nicht. Diese Frage

lässt sich auch mit bedeutend weniger Rechenaufwand beantworten. Simulationen haben gezeigt, dass Datensätze mit einem höheren absoluten Maximum auch im Durchschnitt bei verschiedenen Werten für die Krankheitsmodellparameter höhere LOD-Score-Funktionswerte annehmen. Hat also Datensatz A ein höheres absolutes Maximum als Datensatz B, so wird ein beliebiger zufällig ausgewählter Funktionswert aus A im Mittel größer sein als ein ebenfalls zufällig ausgewählter Funktionswert aus B. Der Mittelwert einer Stichprobe von Funktionswerten aus A bei verschiedenen Parameterwerten wird also höher sein als der Mittelwert einer Stichprobe von B. Der Vergleich dieser Stichproben ist bedeutend weniger rechenzeitaufwendig als eine Maximierung mit dem PGRAD-Verfahren. Bei einer Stichprobe von N_s Werten aus der LOD-Score-Funktion fallen lediglich die Kosten für N_s Funktionsauswertungen sowie deren Addition an. Um N_s Funktionswerte zu erzeugen, muss N_s-mal ein Parametervektor w mittels eines Zufallsgenerators erzeugt werden, vgl. (3.1.2). Man kann dazu z.B. das Startwertverfahren nach Methode 4 benutzen, siehe Kap. 3.3. Das Stichprobenverfahren wird formal wie folgt formuliert:

Bemerkung 4.2.1. Sei $\left\{g_1^A, \ldots, g_{N_s}^A\right\}$ die Stichprobe der MOD-Score-Funktion g^A eines Datensatzes A und $\left\{g_1^B, \ldots, g_{N_s}^B\right\}$ die Stichprobe der MOD-Score-Funktion g^B eines Datensatzes B. Dann ist

$$\max g^B < \max g^A \quad \text{“} \Leftrightarrow \text{“} \quad \frac{1}{N_s}\sum_{i=1}^{N_s} g_i^B < \frac{1}{N_s}\sum_{i=1}^{N_s} g_i^A \quad \Leftrightarrow \quad \sum_{i=1}^{N_s} g_i^B < \sum_{i=1}^{N_s} g_i^A. \tag{4.2.1}$$

Selbstverständlich handelt es sich bei der ersten Äquivalenz nicht um eine mathematisch sichere Aussage, es gibt Gegenbeispiele. Jedoch trifft die Aussage in der Praxis oft genug zu, um von Nutzen zu sein, obwohl eine kleine Fehlerquote bleibt. Allerdings können wir auch mit Hilfe des PGRAD-Verfahrens nie mit Sicherheit sagen, ob wir das absolute Maximum (also den “richtigen“ MOD-Score) und nicht ein lokales, kleineres Maximum gefunden haben.

4.3 Power

Setzen wir ein Testniveau von $\alpha = 5\%$ als Kriterium für eine Entscheidung für die Alternativhypothese, stellt sich die Frage, wie oft sich das Verfahren korrekterwei-

se für die Alternativhypothese (H_1) entscheidet, wenn Kopplung gegeben ist. Die Wahrscheinlichkeit, diese Entscheidung richtig zu treffen, nennt man die *Power* des Verfahrens. Sie ist die Gegenwahrscheinlichkeit zur Wahrscheinlichkeit für einen Fehler 2. Art. Um die Power des Verfahrens zu schätzen, benötigt man Datensätze, bei denen bekannt ist, dass Kopplung vorliegt; idealerweise, weil man sie unter H_1 simuliert hat. Man berechnet dazu für jeden Datensatz den MOD-Score und den p-Wert bzw. nur den p-Wert mittels des PGRAD- oder Stichprobenverfahrens. Anschließend berechnet man den relativen Anteil der Datensätze, die einen p-Wert $\leq 5\%$ besitzen und erhält damit einen Schätzer für die Power des Verfahrens. Formal:

Definition 4.3.1. Sei K die Anzahl der Datensätze, die für die Power-Berechnung benutzt werden. Sei $K_p \leq K$ die Anzahl der Datensätze, deren p-Wert 5% nicht übersteigt. Dann bezeichnen wir mit

$$\text{Pow} := \frac{K_p}{K} \tag{4.3.1}$$

die (geschätzte) Power des Verfahrens für das entsprechende Szenario der Alternativhypothese.

Die Power wird von der Art der Daten, die analysiert werden, entscheidend beeinflusst. Vergleichen wir dazu Bild 2.1. Bei der geringen Standardabweichung lässt sich von einem gegebenen Phänotyp leicht auf den zugrundeliegenden Genotyp schließen: Wählt man einen Punkt auf der x-Achse (Phänotyp) und betrachtet die Wahrscheinlichkeitsdichte der drei Gaußfunktionen an dieser Stelle, so wird höchstens eine der drei Funktionen an dieser Stelle eine Dichte aufweisen, die deutlich größer als Null ist (man beachte, dass die Funktionen in der Grafik auf ganz $\mathbb{R}$ definiert sind, allerdings weit abseits ihrer Mittelwerte so stark abfallen, dass die Funktionswerte nahezu mit der x-Achse zusammenfallen). Dieser Unterschied der Dichten erlaubt bei vorgegebener Genotyp-Phänotyp-Relation einen sicheren Schluss, welche Dichtefunktion bzw. welcher Genotyp dem Phänotyp der Person zugrundeliegt. Der Krankheitsgenotyp einer bestimmten Person ist also leicht rekonstruierbar, was im Falle tatsächlich vorliegender Kopplung zwischen Marker und Krankheitsgenort wiederum zu höheren MOD-Scores und kleineren p-Werten führt. Dies wiederum wird eine hohe Power zur Folge haben. Liegen allerdings die Mittelwerte der Dichtefunktionen nahe beieinander, oder sind ihre Standardabweichungen größer, dann "überlappen" sich die Funktionen und die Zuordnung des Genotyps ist nicht mehr so eindeutig. Gegenüber

ersterem Beispiel ist ein Abfall der Power zu erwarten. Auch die Größe des Stammbaums spielt eine Rolle. Kopplung ist über viele Generationen hinweg mit höherer Sicherheit zu entdecken als über eine Generation, wo sich eine Kosegregation auch rein zufällig ergeben kann. Auch die Anzahl der Familien im untersuchten Datensatz ist wichtig, da sich der Gesamt-LOD-Score aus der Summe der LOD-Scores für eine Familie zusammensetzt. Je mehr Familien der Datensatz beinhaltet, desto höher ist im Allgemeinen der LOD-Score, und damit die Power, falls tatsächlich Kopplung vorliegt.

4.4 Wahrscheinlichkeit für einen Fehler 1. Art

Liegt in den untersuchten Daten keine Kopplung vor, d.h. gilt die Nullhypothese, so kann es zufallsbedingt trotzdem zu einem hohen MOD-Score bzw. einem kleinen p-Wert von unter 5% kommen. In diesem Fall würden wir zu Unrecht für Kopplung entscheiden. Diese Situation nennt man einen *Fehler 1. Art*, die Wahrscheinlichkeit dafür wird mit α bezeichnet:

$$\alpha := P\left(\text{Entscheidung für } H_1 \mid H_0\right). \tag{4.4.1}$$

Für eine Methode wie das PGRAD- oder das Stichprobenverfahren ist es wichtig zu wissen, wie wahrscheinlich es ist, einen solchen Fehler zu begehen, wenn in den Daten tatsächlich keine Kopplung vorliegt. Gehorcht eine Teststatistik (in diesem Falle der MOD-Score) einer stetigen Verteilung, weiß man, dass die p-Werte unter H_0 theoretisch gleichverteilt sind. Die Wahrscheinlichkeit α, entsprechend dem Entscheidungskriterium für H_1, einen p-Wert $\leq 5\%$ zu erhalten und damit einen Fehler 1. Art zu begehen, ist also theoretisch 5%. Da die Verteilung des MOD-Scores im quantitativen Fall jedoch unbekannt ist und es durch numerische Ungenauigkeiten zu Abweichungen kommen kann, muss auch die tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art α empirisch validiert werden. Hierzu bedienen wir uns der gleichen Methode wie bei der empirischen Bestimmung der Power: Wir simulieren Datensätze unter der Nullhypothese und berechnen jeweils den p-Wert. Dann bestimmen wir den relativen Anteil der Nullhypothesenreplikate, deren p-Wert 5% nicht übersteigt. Falls keine Kopplung in den simulierten Datensätzen vorliegt, gilt also mit der Notation der Definition 4.3.1 analog:

$$\alpha_{\text{empirisch}} := \frac{K_p}{K}. \tag{4.4.2}$$

Die Wahrscheinlichkeit für einen Fehler 1. Art wird auch als Signifikanzniveau oder Testniveau bezeichnet.

5 Implementierung und Validierung der Methode

Die in den Kapiteln 2 bis 4 beschriebene Methode wurde im Rahmen dieser Arbeit in das Software-Paket GENEHUNTER implementiert. Hierbei handelt es sich um nicht kommerzielle Software zur Kopplungsanalyse, die ursprünglich von Kruglyak, Daly, Reeve-Daly und Lander entwickelt wurde (Kruglyak *et al.*, 1996). Sie ist seitdem mehrere Male modifiziert und erweitert worden, siehe z.B. Kong & Cox (1997), Strauch *et al.* (2000), Dietter *et al.* (2004), Dietter *et al.* (2007) und Mattheisen *et al.* (2008). Die Erweiterung mit der hier vorgestellten Methode für quantitative Phänotypen nennt sich GENEHUNTER-QMOD. Sie ist im Internet unter *http://www.helmholtz-muenchen.de/genepi/downloads* frei erhältlich.

5.1 Vergleich mit anderen Verfahren

Die hier entwickelte und vorgestellte Methode zur Kopplungsanalyse ist nicht das einzige verfügbare Verfahren. Diverse andere Methoden sind für diese Problemstellung ebenfalls geeignet und wurden bereits in verschiedene Softwarepakete implementiert. Neben GENEHUNTER (Kruglyak *et al.*, 1996) sind z. B. MERLIN (Abecasis *et al.*, 2002) und ALLEGRO (Gudbjartsson *et al.*, 2000) zwei weitere wichtige Programme, mit denen sich Kopplungsanalyse durchführen lässt. Ein Überblick über die gebräuchlichste Kopplungsanalyse-Software findet sich in Dudbridge (2003). Eine Methode, die sich mit quantitativen Phänotypen befasst, ist die *Varianzkomponentenanalyse*. Sie ist in diverse Softwarepakete implementiert worden, z.B. GENEHUNTER oder SOLAR (Almasy & Blangero, 1998). Bei diesem Verfahren werden zusätzlich zum LOD-Score weitere Informationen bezüglich der Varianzkomponenten des Phänotyps geliefert. Ein weiteres Verfahren, um quantitative Phänotypen zu untersuchen, ist die *Haseman-Elston-Regression* (Haseman & Elston, 1972). Allerdings nutzt die

Haseman-Elston-Regression bei der Analyse nur Verwandtschaftsbeziehungen zwischen Geschwisterpaaren aus und liefert keine spezifischen Informationen über die Genotyp-Phänotyp-Relation. Das MDE-Verfahren (Ziegler & Kastner, 1997), eine weiterentwickelte Methode, die auf der Haseman-Elston-Regression basiert, nutzt hingegen Paare von Verwandten verschiedener Art bei der Analyse. Wie die Varianzkomponentenanalyse setzt sie jedoch gleiche Restvarianzen für alle Krankheitsgenotypen voraus. Parametrische Kopplungsanalyse für quantitative Phänotypen wurde in die Softwarepakete LINKAGE/FASTLINK (Lathrop & Lalouel, 1984), (Cottingham *et al.*, 1993) und PAP (Hasstedt & Cartwright, 1981) implementiert. LINKAGE setzt ebenfalls gleiche Restvarianzen voraus, mit PAP lassen sich auch genotypspezifische Restvarianzen modellieren. Beide Programme benutzen jedoch den Elston-Stewart-Algorithmus (Elston & Stewart, 1971), der die Analyse auf eine kleine Anzahl an Markern beschränkt. Dies ist unvorteilhaft für eine Anwendung bei Genkartierungsprojekten, die SNPs verwenden. Letztere sind weniger informativ als Mikrosatelliten-Marker, so dass eine größere Anzahl an SNPs zusammen analysiert werden muss.

Im Folgenden werden zwei wichtige Verfahren genauer vorgestellt, mit denen wir das PGRAD- und Stichprobenverfahren vergleichen wollen, nämlich die Varianzkomponentenanalyse und die Haseman-Elston-Regression.

5.1.1 Varianzkomponentenanalyse

Das Modell, das der Varianzkomponentenanalyse zugrunde liegt, ist das Folgende: Einen quantitativen, stetigen Phänotyp P kann man in der Gesamtpopulation als normalverteilte Zufallsvariable auffassen. Bei jeder Person setzt sich der Phänotyp aus einem Anteil G zusammen, der durch die Gene der Person erzeugt wird, sowie einem Anteil E, der durch äußere Umwelteinflüsse bestimmt wird. Es gilt also

$$P = G + E. \tag{5.1.1}$$

G und E kann man wieder als normalverteilte Zufallsvariablen auffassen, mit jeweils eigenen Erwartungswerten und Varianzen. Unter Annahme der Unabhängigkeit von G und E setzt sich die Varianz des Phänotyps P aus der Varianz von G und E zusammen:

$$V_P = V_G + V_E. \tag{5.1.2}$$

Bei einem einzelnen Genort mit zwei Allelen (A_1 und A_2) führt man dabei zwei weitere Variablen für den Effekt der drei möglichen Genotypen ein: a ist der halbe Abstand zwischen den Mittelwerten des Phänotyps für die beiden homozygoten Genotypen. Man definiert dabei die Mitte zwischen den beiden Genotypen als 0. Dann ist der mittlere phänotypische Effekt des Genotyps $A_1A_1 = -a$ und des Genotyps $A_2A_2 = a$. Die Variable d gibt, ausgehend von 0, den Mittelwert für den heterozygoten Genotyp an. Für $d = a$ oder $d = -a$ hat man einen vollständig dominanten oder rezessiven Erbgang. Deswegen heißt d auch Dominanzeffekt und a additiver Effekt. Die genetische Varianz V_G wird entsprechend in eine additive und eine dominante Komponente zerlegt:

$$V_G = V_A + V_D. \tag{5.1.3}$$

Auf ähnliche Weise lässt sich die Varianz der Umwelteinflüsse V_E in eine Varianz V_S, die individuelle spezifische Umwelteinflüsse berücksichtigt, sowie in eine Varianz V_C, die auf alle Mitglieder einer bestimmten Familie wirkende Einflüsse modelliert, zerlegen:

$$V_E = V_S + V_C. \tag{5.1.4}$$

Mit (5.1.3) und (5.1.4) wird (5.1.2) zu

$$V_P = V_A + V_D + V_S + V_C. \tag{5.1.5}$$

Dies sind die Parameter, die mit der Varianzkomponentenanalyse geschätzt werden sollen.

Letztendlich wird im Rahmen eines Likelihoodquotiententests wieder ein LOD-Score aus dem Datensatz berechnet und über die Varianzparameter maximiert. Die Werte, die den LOD-Score optimieren, sind die Maximum-Likelihood-Schätzer für die Parameter, die die beobachteten Daten am besten erklären. Bei der Optimierung wird Fishers Scoring-Verfahren benutzt. Die p-Wert-Bestimmung zu einem LOD-Score ist bei der VC-Analyse glücklicherweise nicht allzu aufwendig. Man kann zeigen, dass der LOD-Score unter bestimmten Bedingungen einer modifizierten χ^2-Verteilung folgt, vgl. z. B. Pratt *et al.* (2000). Damit lässt sich der p-Wert ohne rechenintensive Simulationen bestimmen, es genügt, das Integral unter der Verteilungsdichte zu kennen. Beim Likelihoodquotiententest sind die unter der Nullhypothese keiner Kopplung

angenommenen Varianzen für die genetische additive und dominante Komponente V_A und V_D jeweils 0. Es existieren also im Nenner zwei freie Parameter weniger als im Zähler, der über beide Varianzkomponenten maximiert wird. Man kann zeigen, dass dann der mit $2\ln(10) \approx 4{,}6$ multiplizierte LOD-Score der Varianzkomponentenanalyse eine Zufallsvariable mit asymptotischer Verteilungsfunktion $\frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$ ist, vgl. Rowe (2008), S.79. χ_0^2, χ_1^2 und χ_2^2 stehen dabei für eine Punktmasse bei 0 bzw. für die χ^2-Verteilungsfunktionen mit entsprechenden Freiheitsgraden.

Eine ausführliche Beschreibung der Varianzkomponentenanalyse findet sich z. B. in Camp & Cox (2002), Kapitel 4.

5.1.2 Haseman-Elston-Regression

Ein anderes, gebräuchliches Verfahren zur Kopplungsanalyse von quantitativen Phänotypen ist die *Haseman-Elston-Regression*. Es handelt sich dabei um ein nichtparametrisches Verfahren und ist nur auf Geschwisterpaare anwendbar. Es beruht auf der Überlegung, dass sich Geschwister, die sich genotypisch ähnlich sind, auch phänotypisch ähneln sollten. Andersherum sollten dann genotypisch unähnliche Geschwister auch abweichende Phänotypen haben. Ein Maß für die genotypische und phänotypische Ähnlichkeit wird dabei wie folgt definiert: Seien y_1 und y_2 die Phänotypen der beiden Geschwister, dann wird die quadrierte phänotypische Differenz $(y_1 - y_2)^2$ als Maß für den Abstand der beiden Werte benutzt. Die Ähnlichkeit der Genotypen am untersuchten Markerlocus wird über die Anzahl der Allele, die die Geschwister "identical by descent" (ibd) gemeinsam haben, gemessen. Diese Variable kann die Ausprägungen $0, 1, 2$ annehmen. Diese beiden Größen sollen nun durch eine lineare Regression miteinander in Zusammenhang gebracht werden, d. h. man legt eine Ausgleichsgerade durch die Datenpunkte für alle Geschwisterpaare, die diese möglichst gut approximiert. Beeinflusst der untersuchte Locus den quantitativen Phänotyp, wird mit zunehmender Anzahl an ibd-Allelen die quadrierte phänotypische Differenz abnehmen. Man erwartet also eine Regressionsgerade mit negativer Steigung. Für eine ausführlichere Beschreibung des Verfahrens und seiner Eigenschaften wird z.B. auf Ziegler (1999) verwiesen, oder auf die bereits erwähnte Originalarbeit Haseman & Elston (1972). Das Ergebnis der Berechnung der Haseman-Elston-Regression ist u. a. der sogenannte *t-Score*. Er ist t-verteilt, die Anzahl der Freiheitsgrade ist die Anzahl der Geschwisterpaare minus 1 (Franke *et al.*, 2005). Auf diese Weise läßt sich zu einem Datensatz bzw. zu einem Score leicht der p-Wert bestimmen.

5.2 Datensimulation

Um das tatsächliche Signifikanzniveau und die Power einer Methode messen zu können, benötigt man Daten, von denen man bereits weiß, ob ihnen ein Szenario mit Kopplung zugrunde liegt oder nicht, d.h. ob H_0 oder H_1 gilt. Will man weiterhin wissen, wie gut die Methode die genotypspezifischen Phänotyp-Parameter schätzt, ist die Kenntnis des tatsächlichen Krankheitsmodells vonnöten. Dazu ist es erforderlich, Daten unter der entsprechenden Hypothese bzw. dem jeweiligen Modell zu simulieren. Für GENEHUNTER und seine Erweiterungen müssen die Datensätze im sogenannten LINKAGE-Format vorliegen. Es handelt sich um ein Datenformat, das zwei Textdateien beinhaltet. Informationen über die beobachteten genetischen Marker und den Krankheitsphänotyp sowie die Familienstruktur der Personen im Datensatz werden in einem sog. *pedigree*-file übergeben. Die Rekombinationsfrequenzen zwischen den einzelnen Markern sowie die Allelfrequenzen der einzelnen Allele auf den entsprechenden Markern finden sich in einem sog. *data*-file wieder.

Für die Simulationen in dieser Arbeit wurde das Programm SIBSIM von Daniel Franke benutzt (Franke *et al.*, 2006). Es ist wie GENEHUNTER frei erhältlich. SIBSIM erstellt simulierte Datensätze für quantitative Phänotypen in Form von pedigree- und data-files im LINKAGE-Format. Die Anzahl der Familien, die Familienstruktur sowie Anzahl der Marker, ihre Position, Zahl der Allele und ihre Frequenzen sind dabei frei wählbar. Auch die Position des Krankheitslocus ist beliebig wählbar, man kann also Szenarien sowohl unter der Alternativ- als auch unter der Nullhypothese simulieren. Auch die Parameter der Genotyp-Phänotyp-Relation sind beeinflussbar. Allerdings war es zunächst nicht möglich, die genotypspezifischen Erwartungswerte μ_i unabhängig voneinander zu wählen. Auch verschiedene Varianzen σ_i^2 können nicht gewählt werden. Die Software wurde vom Autor insoweit verändert, als dass man nun die Erwartungswerte der Dichtefunktionen der einzelnen Genotypen direkt und unabhängig voneinander setzen kann.

Für komplexere Szenarien mit beispielsweise genotypspezifisch nicht normalverteilten Phänotypen oder Datensätzen mit nicht zufälligem Rekrutierungsschema wurden verschiedene selbst erstellte Skripte für das Statistik-Programmpaket R (R Development Core Team, 2008) benutzt, um die mit SIBSIM erstellten pedigree files nachträglich zu modifizieren. Auch genotypspezifisch verschiedene Varianzen konnten so berücksichtigt werden.

6 Ergebnisse

In diesem Kapitel werden wir die Power sowie die tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art unseres Verfahrens unter diversen Bedingungen testen und mit zwei bestehenden Verfahren, der Varianzkomponentenanalyse und der Haseman-Elston-Regression, vergleichen. Da das PGRAD-Verfahren neben dem Test auf Kopplung weitere Informationen über die Genotyp-Phänotyp-Relation liefert, werden wir weiterhin die Schätzung der Parameter μ_i und σ_i untersuchen. Zum Schluss werden mit dem Verfahren reale Daten untersucht. Als Beispiel dient ein Datensatz über die genetisch bedingte Allergisierung auf Hausstaubmilben.

6.1 PGRAD- und Stichprobenverfahren

Ein erstes Szenario, mit dem die Power des PGRAD-Verfahrens berechnet wurde, besteht aus 200 simulierten Datensätzen. Jeder enthält 300 Familien mit zwei Elternteilen und zwei Kindern (sib-pair-Szenario). Es wurde ein additives Krankheitsmodell benutzt, mit den Parametern $\mu_{(+,+)} = 20{,}0$, $\mu_{(m,+)} = 40{,}0$, $\mu_{(m,m)} = 60{,}0$ für die Mittelwerte der genotypspezifischen Dichtefunktionen, sowie $\sigma^2_{(+,+)} = \sigma^2_{(m,+)} = \sigma^2_{(m,m)} = 100{,}0$ für die Varianzen, vgl. Definition 2.1.1. Bei gleichen Varianzen bezeichnen wir σ_i^2 auch als *error-effect*. Zusätzlich wurde ein familienspezifischer Störterm simuliert und zum Phänotyp eines jeden Familienmitgliedes addiert. Bei diesem familienspezifischen Störterm handelt es sich um eine normalverteilte Zufallsvariable mit Erwartungswert 0 und Varianz 70. In Anlehnung an die Notation in SIBSIM bezeichnen wir im Folgenden die Varianz des Störterms als *family-effect*. Zusätzlich wurde ein vollständig informativer Marker simuliert, der direkt auf den Krankheitslocus platziert wurde, entsprechend vollständiger Kopplung zwischen dem Marker und dem Krankheitslocus ($\theta = 0$). Um den p-Wert eines jeden Datensatzes zu berechnen, wurden für Szenario 1 zu jedem unter der H_1-Hypothese (Kopplung) erstellten Datensatz mit GENEHUNTER-QMOD 5.000 Replikate unter der Nullhypothese (im Folgenden

H_0-Replikate genannt) erstellt und jeweils der MOD-Score berechnet. Hierbei wurde das PGRAD -Verfahren mit je 10 Startwerten benutzt.

In einem zweiten Szenario wurde ein stärkerer *error-effect* von 150 sowie ein ebenfalls größerer *family-effect* von 100 gewählt. Die restlichen Eigenschaften (Stammbaumstruktur, Anzahl der Datensätze und Familien, Marker) blieben gegenüber dem ersten Szenario unverändert. Für das zweite Szenario wurde die Stichproben-Methode benutzt. Aufgrund der schnelleren Berechnung war es möglich, pro Datensatz 50.000 Nullhypothesenreplikate zu berechnen.

Die sich ergebende Power sowie alle anderen relevanten Daten der Szenarien 1 und 2 sind in Tabelle 6.1 zusammengefasst. Das Signifikanzniveau α betrug jeweils 5%.

Tabelle 6.1: PGRAD- und Stichprobenverfahren

	Szenario 1			Szenario 2		
Zahl der Datensätze unter H_1			200			200
Zahl der Familien			300			300
Stammbaumstruktur			sib-pair-Familie			sib-pair-Familie
H_0-Replikate pro H_1-Datensatz			5.000			50.000
$\mu_{(+,+)}$, $\mu_{(m,+)}$, $\mu_{(m,m)}$	20,0	40,0	60,0	20,0	40,0	60,0
Varianz *error-effect*			100,0			150,0
Varianz *family-effect*			70,0			100,0
Zahl der Marker			1			1
Allele pro Marker			20			20
Optimierungsverfahren			PGRAD			Stichproben
Anzahl der Startwerte			10			-
max. Anzahl Iterationen in PGRAD			500			-
Power			78,0%			84,5%

6.2 Power und Signifikanzniveau

Da sich das Stichprobenverfahren aufgrund der höheren Rechengeschwindigkeit besser zur Bestimmung des p-Wertes und damit auch der Power und des tatsächlichen Signifikanzniveaus eignet, wird es im Folgenden für alle Power- und Signifikanzniveau-Analysen verwendet.

Wir werden zunächst das Stichprobenverfahren mit Szenarien mit zufällig rekrutierten Familien untersuchen und mit der Varianzkomponentenanalyse und der

Haseman-Elston-Regression vergleichen. Bei diesen Szenarien handelt es sich um Daten, in denen sich die Genotypen der Founder so zusammensetzen, wie man sie auch in der Grundpopulation antreffen würde. Man kann sich vorstellen, dass man die einzelnen Familien zufällig und nach keinem speziellen Auswahlkriterium aus der Bevölkerung auswählt und für die Analyse verwendet.

Bis auf eine Ausnahme wurde dabei weiterhin das sib-pair-Szenario benutzt. Auch die Anzahl der Familien in einem Replikat, Anzahl und Eigenschaften der Marker sowie das Signifikanzniveau von 5% wurden gegenüber Szenario 1 und 2 nicht verändert.

In Szenario 3 wurden wieder 200 Datensätze konstruiert. Um den p-Wert zu erhalten, verarbeitete das Stichprobenverfahren für jeden Datensatz 5000 Nullhypothesenreplikate. Für den Phänotyp wurde ein additives Modell ohne Dominanzeffekte und mit gleichen Varianzen gewählt. Ein zusätzlicher familienspezifischer Störterm ist nicht im Modell enthalten. Die Power wurde immer unter den Analyseoptionen berechnet, die dem simulierten Modell entsprechen, in diesem Fall keine Dominanzeffekte und gleiche Varianzen. Auch in der Varianzkomponentenanalyse (im Folgenden entsprechend der englischen Bezeichnung *variance components analysis* auch mit *VCA* abgekürzt) wurden, soweit möglich, die Optionen dem tatsächlichen Krankheitsmodell angepasst. In diesem Szenario liefert das Stichprobenverfahren mit 82% eine akzeptable Power, liegt jedoch sowohl hinter der VCA als auch der Haseman-Elston-Regression zurück (Power von 98% bzw. 91,5%, siehe Tabelle 6.2). Ein Histogramm der p-Wert-Verteilung der Berechnung mit dem Stichprobenverfahren findet sich in Bild 6.1.

Da Stammbäume mit zusätzlichen Familienmitgliedern auch zusätzliche Informationen für die Kopplungsanalyse liefern, ist bei gleichbleibender Zahl der Familien im Datensatz ein Anstieg der Power zu erwarten, wenn man komplexere Stammbäume analysiert. Daher bestand das nächste Szenario, das wir untersuchten, aus Familien, die aus den beiden Foundern und vier Kindern bestehen. Belässt man die Genotyp-Phänotyp-Relation wie in Szenario 3, erhöht sich die Power des Stichprobenverfahrens und der VCA auf 100%. Da dies keinen aussagekräftigen Vergleich gestattet, erhöhten wir den Störterm des Phänotyps, d.h. die Varianzen σ_i^2, auf 350,0. Die Ergebnisse der Poweranalyse finden sich unter Szenario 4, ebenfalls in Tabelle 6.2. Wie bei Szenario 3 liegt die Power des Stichprobenverfahrens bei 82%. Die Varianzkomponentenanalyse und die Haseman-Elston-Regression fallen jedoch deutlich

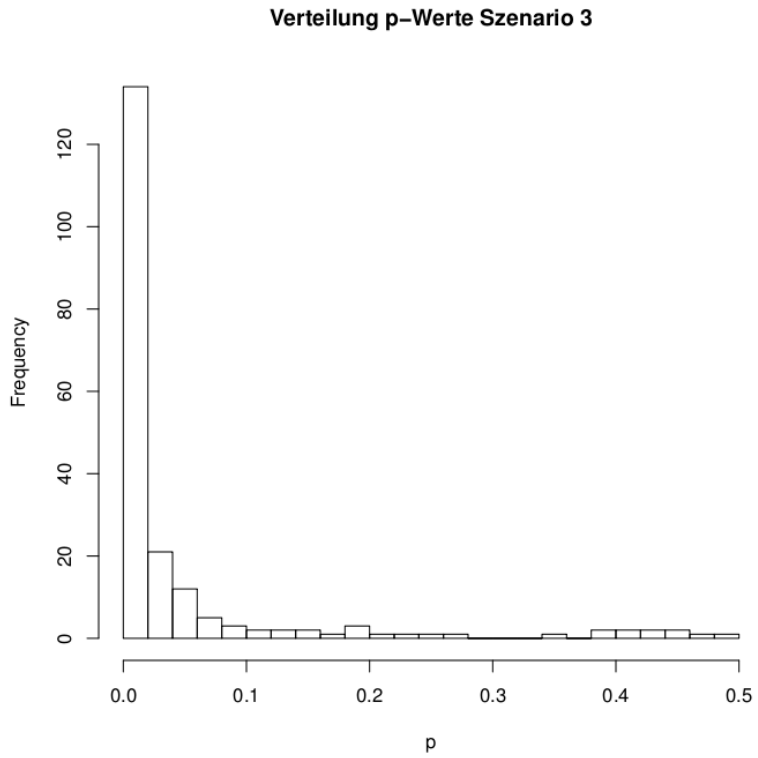

Bild 6.1: Histogramm der p-Wert-Verteilung aus Szenario 3

ab und schneiden mit einer Power von 77% bzw. 22,5% nun schlechter ab als das Stichprobenverfahren.

Da GENEHUNTER-QMOD unterschiedliche Varianzen σ_i^2 modellieren kann, wurde bei den Simulationen in Szenario 5 angenommen, dass sich die genotypspezifische Varianz mit zunehmender Anzahl mutanter Allele m erhöht. Die restlichen Parameter blieben wie in Szenario 3. Hier verzeichnet die VCA nur noch einen Powergewinn von 7,6 % gegenüber GENEHUNTER-QMOD. Man beachte, dass es in der VCA nicht möglich ist, ungleiche Varianzen im Modell zu berücksichtigen. Die VCA trifft also eine falsche Annahme. Die Haseman-Elston-Regression verzeichnet hier eine Power von 74,5%; diese ist damit niedriger als die des Stichprobenverfahrens und der VCA.

Es kann nicht bei jedem Phänotyp davon ausgegangen werden, dass die genotypspezifischen Verteilungen, wie in Bild 2.1 dargestellt, einer Normalverteilung entsprechen. Daher wurden in Szenario 6 Daten simuliert, in denen der Phänotyp nicht normalverteilt ist. Die Normalverteilungen wurden für jeden Genotyp durch Lognormalverteilungen ersetzt:

$$f_i(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_i y} \exp\left(-\frac{(\ln y - \mu_i)^2}{2\sigma_i^2}\right) & y > 0 \\ 0 & y \leq 0 \end{cases} \tag{6.2.1}$$

Jeder Genotyp hat nun eine spezifische Lognormalverteilung mit Erwartungswert μ_i und Varianz σ_i^2. Die restlichen Bedingungen von Szenario 6 (Anzahl Datensätze, Familienstruktur, Marker, Anzahl Nullhypothesenreplikate) wurden gegenüber den Szenarien 3 und 5 nicht verändert. Die genauen Parameter der Lognormalverteilungen sowie die Power der drei Verfahren sind wieder Tabelle 6.2 zu entnehmen. Hier liefert nun GENEHUNTER-QMOD eine deutlich höhere Power als die VCA, deren Power um exakt 35% geringer ist. Noch schlechter kommt die Haseman-Elston-Regression mit lognormalverteilten Daten zurecht: Die Power beträgt hier 13%, was für einen statistischen Test inakzeptabel ist.

Tabelle 6.2: Power-Vergleich des Stichprobenverfahrens, der VCA und der Haseman-Elston-Regression

	$\mu_{(+,+)}$	$\mu_{(m,+)}$	$\mu_{(m,m)}$	$\sigma^2_{(+,+)}$	$\sigma^2_{(m,+)}$	$\sigma^2_{(m,m)}$	P_S	P_{VCA}	P_{HE}
Szenario 3	20,0	40,0	60,0	200,0	200,0	200,0	0,82	0,98	0,915
Szenario 4	20,0	40,0	60,0	350,0	350,0	350,0	0,82	0,77	0,225
Szenario 5	20,0	40,0	60,0	100,0	225,0	400,0	0,874	0,95	0,745
Szenario 6	20,0	100,0	180,0	2,25	6,25	9,0	0,645	0,295	0,13

P_S steht für die Power des Stichprobenverfahrens, P_{VCA} ist die Power der Varianzkomponentenanalyse und P_{HE} jene der Haseman-Elston-Regression.

Zur Validierung des Signifikanzniveaus wurden ebenfalls die Szenarien 3, 4, 5 und 6 benutzt. Alle Parameter blieben gleich, nur der Krankheitslocus wurde nicht direkt an der Position des Markers simuliert, sondern in einem Abstand von 1000 cM, was einer Rekombinationsfrequenz entspricht, die um weniger als 10^{-90} von $\frac{1}{2}$ abweicht (Haldane-Mapping-Funktion). Die tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art findet sich in Tabelle 6.3. Das Stichprobenverfahren ist tendenziell etwas zu konservativ, wobei die Abweichung des Signifikanzniveaus vom theoretischen Wert von 5% im Szenario 4 mit erweiterten Stammbäumen am kleinsten ist. Phänotypen mit verschiedenen genotypspezifischen Varianzen führen zur größten (konservativen) Abweichung. Lediglich in Szenario 3 zeigt sich eine leichte antikonservative Tendenz. Die Varianzkomponentenanalyse hält das Niveau in den Szenarien 3, 4 und 5 relativ gut ein, ist aber in Szenario 6 (nicht normalverteilte Daten) extrem liberal. Man

beachte, dass in diesem Fall die Power der VCA sehr niedrig ist. Die Haseman-Elston-Regression hält in allen Szenarien das Fehlerniveau relativ gut ein. In den Szenarien mit nicht normalverteilten Phänotypen oder verschiedenen Varianzen ist sie allerdings wie das Stichprobenverfahren leicht konservativ.

Tabelle 6.3: Tatsächlicher Fehler 1. Art des Stichprobenverfahrens, der VCA und der Haseman-Elston-Regression

	α_S	α_{VCA}	α_{HE}
Szenario 3	0,06	0,0525	0,05
Szenario 4	0,046	0,042	0,042
Szenario 5	0,024	0,04	0,038
Szenario 6	0,036	0,134	0,034

α_S ist der tatsächliche Fehler 1. Art des Stichprobenverfahrens, α_{VCA} jener der Varianzkomponentenanalyse und α_{HE} der Haseman-Elston-Regression. Die Parameter der Szenarien finden sich in Tabelle 6.2.

6.3 Nicht-zufälliges Ascertainment

Da die Auswahl von Familien für eine Kopplungsanalyse in der Praxis oftmals nicht zufällig aus der Population, sondern nach gewissen Kriterien erfolgt, untersuchen wir auch Szenarien, in denen die Daten nach einem bestimmten Rekrutierungsschema, engl. *ascertainment criterion*, erhoben werden. Ein solches Rekrutierungsschema kann z.B. in der Bedingung bestehen, dass eine Familie mindestens ein Kind mit einem bestimmten Phänotyp enthalten muss. Für die Poweranalyse wurden wieder 200 Datensätze (SIBSIM-Replikate) erzeugt. Die Familienstruktur (sib-pair-Szenario), Marker und Krankheitslocus sowie die Anzahl der Nullhypothesenreplikate zur Bestimmung des empirischen p-Wertes blieben wie in den Szenarien 3, 5 und 6. Zur Bestimmung des tatsächlichen Signifikanzniveaus wurden wiederum 500 Datensätze benutzt. Bei der Simulation der Daten mit SIBSIM wurden hier jedoch nur spezielle Familien ausgewählt: Szenario 7 enthält nur Familien mit mindestens einem Kind im höchsten Quartil der Grundpopulation (single proband selection). In Szenario 8 wurde eine Familie nur ausgewählt, wenn entweder beide Kinder im höchsten Quartil der Verteilung oder beide Kinder im niedrigsten Quartil der Verteilung lagen. Szenario 9 besteht aus Familien mit einem Kind im höchsten und dem anderen Kind im

niedrigsten Quartil der Verteilung. Die Bedingungen in Szenario 8 und 9 werden als double proband selection bezeichnet, siehe auch Kleensang *et al.* (2010).

Es wurden Familien simuliert und ausgewählt, bis jeder Datensatz 300 Familien mit der entsprechenden Bedingung enthielt. Die Erwartungswerte der Szenarien 7, 8 und 9 wurden bei den Simulationen auf $\mu_{(+,+)} = 20{,}0$, $\mu_{(m,+)} = 40{,}0$, $\mu_{(m,m)} = 60{,}0$ gesetzt, und die Varianzen auf $\sigma^2_{(+,+)} = \sigma^2_{(m,+)} = \sigma^2_{(m,m)} = 200{,}0$. Die Krankheitsallelfrequenz in der Grundpopulation, aus der die Familien gezogen wurden, betrug 0,1. Die Ergebnisse sind in Tabelle 6.4 zusammengestellt.

Tabelle 6.4: Tatsächlicher Fehler 1. Art und Power des Stichprobenverfahrens, der VCA und der Haseman-Elston-Regression

	α_S	α_{VCA}	α_{HE}	P_S	P_{VCA}	P_{HE}
Szenario 7	0,066	0,0	0,06	0,895	0,04	0,585
Szenario 8	0,032	0,088	0,052	0,94	0,89	0,030
Szenario 9	0,042	0,0	0,052	0,975	0,0	0,495

α_S bezeichnet das Signifikanzniveau des Stichprobenverfahrens, α_{VCA} das Niveau der VCA und α_{HE} jenes der Haseman-Elston-Regression. P_S, P_{VCA} und P_{HE} bezeichnen die Power des jeweiligen Verfahrens.

Diese Untersuchung zeigt, dass sich das Stichprobenverfahren etwas zu liberal unter single proband selection verhält (Szenario 7), und einigermaßen konservativ unter double proband selection (Szenario 8 und 9) ist.

Die VCA ist extrem konservativ unter single proband selection und double proband selection mit beiden Kindern in verschiedenen Quartilen. Allerdings ist sie zu liberal unter double proband selection mit beiden Kindern im gleichen Quartil. Die Haseman-Elston-Regression hingegen hält in allen Rekrutierungsszenarien das Niveau relativ gut ein. Einzig unter single selection ist das Verfahren ein wenig zu liberal. Das Stichprobenverfahren zeigt eine gute Power in allen Rekrutierungsszenarien. Es ist besonders stark unter double proband selection, während es hier zugleich unter der Nullhypothese keiner Kopplung konservativ ist.

Die Power der VCA ist nahe null unter single proband selection und double proband selection mit Kindern in verschiedenen Quartilen (Szenarien 7 und 9). Es sollte erwähnt werden, dass die VCA in diesen Szenarien unter H_0 auch extrem konservativ ist. Allerdings liefert das Verfahren eine gute Power für double proband selection mit Kindern in gleichen Quartilen (Szenario 8), auch wenn es nicht ganz die Power des

Stichprobenverfahrens erreicht und außerdem wie gesagt zu liberal ist.

Die Power der Haseman-Elston-Regression verhält sich genau gegensätzlich zur Power der VCA: Während es in Szenario 7 und 9 noch eine moderate Power zeigt, ist diese mit $0,03$ viel zu niedrig in Szenario 8. Allerdings erreicht die Haseman-Elston-Regression in keinem Szenario auch nur annähernd die Power des Stichprobenverfahrens.

6.4 Parameterschätzung

Wie bereits erwähnt, sucht das hier entwickelte, in GENEHUNTER-QMOD implementierte Verfahren nicht nur nach dem Locus des krankheitsverursachenden Gens, sondern schätzt ebenfalls Parameter, die Informationen über den quantitativen Phänotyp liefern. Hier soll untersucht werden, wie gut diese Parameterschätzung in GENEHUNTER-QMOD funktioniert. Dazu wurden von Szenario 3 und Szenario 8

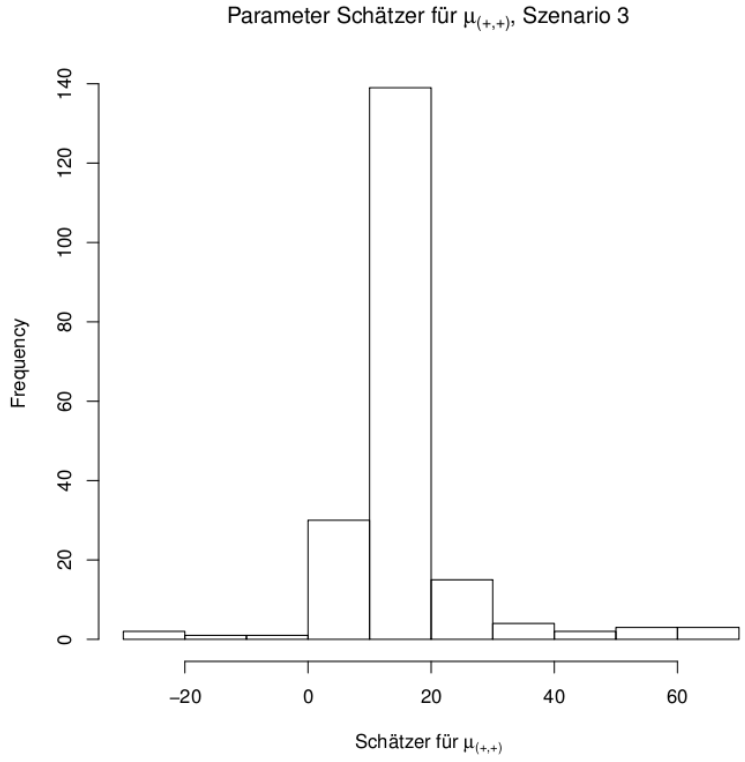

Bild 6.2: Histogramm von 200 Parameterschätzungen von $\mu_{(+,+)} = 20{,}0$ aus Szenario 3

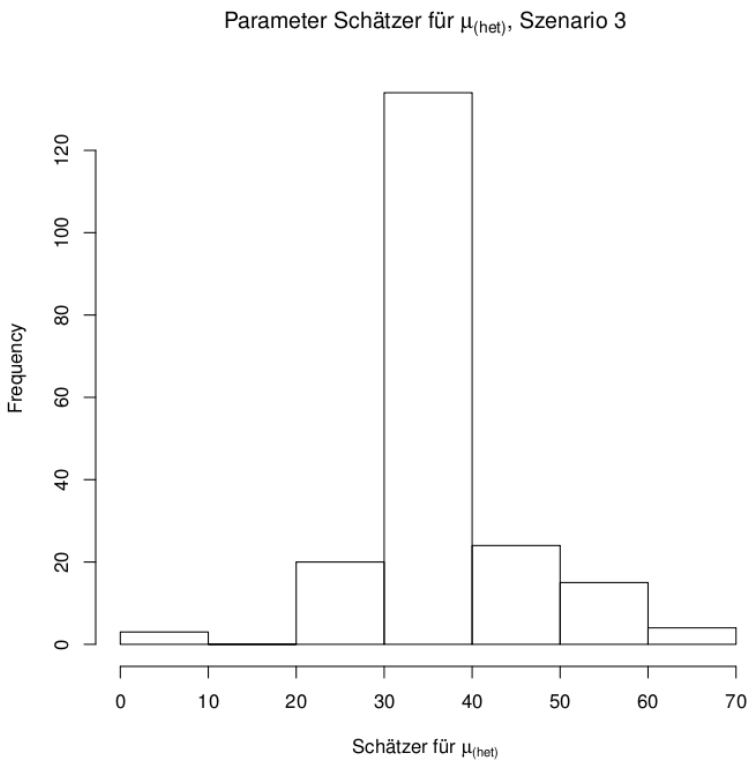

Bild 6.3: Histogramm von 200 Parameterschätzungen von $\mu_{(m,+)} = 40{,}0$ aus Szenario 3

(unter der Alternativhypothese Kopplung) alle 200 Datensätze mit dem PGRAD-Verfahren analysiert und jeweils die Parameterkombination
$w = \left(\mu_{(+,+)}, \mu_{(m,+)}, \mu_{(m,m)}, \sigma_{(+,+)}, \sigma_{(m,+)}, \sigma_{(m,m)}\right)$ zum maximierten LOD-Score ausgegeben. Jeder Parameter des Phänotyps wurde also 200 mal geschätzt. Aus diesen 200 Schätzern wurde jeweils der empirische Mittelwert und die zugehörige Standardabweichung berechnet. Des weiteren wurde der Fehler zwischen dem wahren Parameterwert und dem Mittelwert der Schätzer berechnet. Hierbei ist zu beachten, dass der absolute Fehler zwischen dem Schätzer-Mittelwert und dem wahren Wert wenig Sinn macht, da er kein sinnvolles Maß für die Genauigkeit des Schätzers liefert. Auch der relative Fehler (wahrer Wert - geschätzter Wert)/wahrer Wert ist für die Erwartungswerte μ_i kein gutes Maß, da er bei großen Phänotyp-Werten automatisch geringer ausfällt als bei kleinen Werten. Entscheidend zur Unterscheidung der Phänotypen ist lediglich der Abstand zwischen $\mu_{(+,+)}$, $\mu_{(m,+)}$ und $\mu_{(m,m)}$. Der Fehler eines Erwartungswertes μ_i wird deswegen auf den Abstand von μ_i und

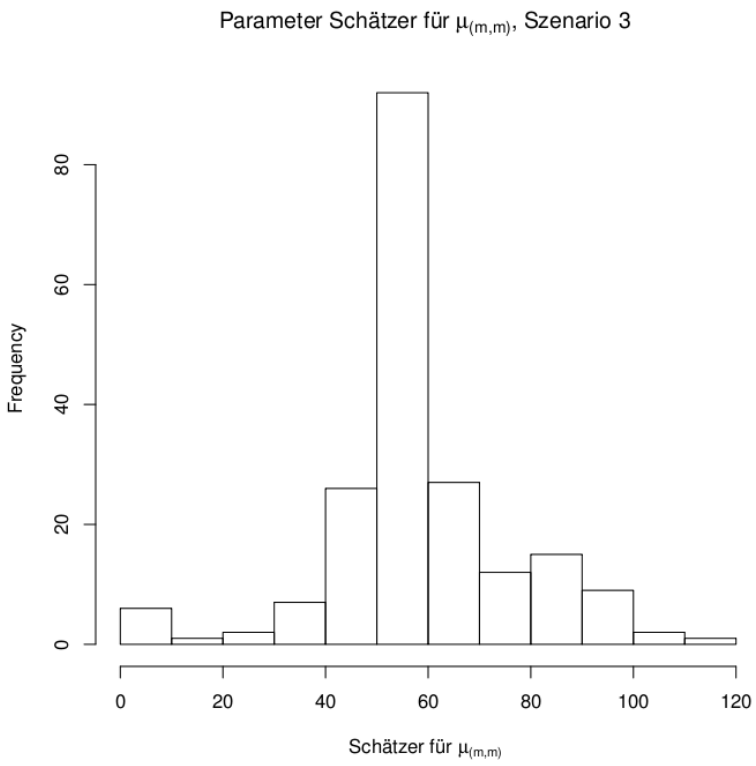

Bild 6.4: Histogramm von 200 Parameterschätzungen von $\mu_{(m,m)} = 60{,}0$ aus Szenario 3

des ihm am nächsten liegenden Erwartungswertes $\mu_{i'}$ normiert. Der Fehler für die Standardabweichungen ist der übliche relative Fehler.

Der Mittelwert und die Standardabweichung für jede Variable sowie der entsprechende Fehler sind in Tabelle 6.5 gegeben.

Um ein besseres Bild von der Verteilung der Schätzer zu bekommen, sind beispielhaft für Szenario 3 in den Bildern 6.2, 6.3 und 6.4 Histogramme der Parameterverteilungen für $\mu_{(+,+)}$, $\mu_{(m,+)}$ und $\mu_{(m,m)}$ gegeben.

Tabelle 6.5: Parameterschätzung für Szenario 3 und 8

	Parameter	Mittelwert	wahrer Wert	Std. Abweichung	Fehler in %
Szenario 3	$\mu_{(+,+)}$	20,0	15,1	10,7	24,5
	$\mu_{(m,+)}$	40,0	37,4	8,8	13,0
	$\mu_{(m,m)}$	60,0	59,6	17,6	2,0
	σ	14,1	7,7	2,74	45,4
Szenario 8	$\mu_{(+,+)}$	20,0	12,8	7,0	36,0
	$\mu_{(m,+)}$	40,0	36,5	8,2	17,5
	$\mu_{(m,m)}$	60,0	60,3	16,5	1,5
	σ	14,1	5,4	3,24	61,7

σ steht für die restliche Standardabweichung in Szenario 3 und 8, die für alle drei Genotypen identisch ist; $\sigma := \sigma_{(+,+)} = \sigma_{(m,+)} = \sigma_{(m,m)}$. Die letzte Spalte beschreibt für die Mittelwerte μ_i die relative Abweichung bezogen auf den Abstand von μ_i und den nächstliegenden Erwartungswert $\mu_{i'}$.

6.5 Anwendung: Hausstaubmilbenallergie

An der Entstehung von Allergien beim Menschen sind mutmaßlich mehrere Gene beteiligt. Die Sensibilisierung auf Allergene bezeichnet man auch als Atopie. Bei einer allergischen Reaktion ruft eine bestimmte Substanz, auch Allergen genannt, eine Überempfindlichkeitsreaktion des menschlichen Immunsystems hervor. Dieses bildet IgE-Antikörper, wenn es mit dem Allergen in Berührung kommt. Meist handelt es sich bei Allergenen um Eiweiße oder Eiweißverbindungen. Die Konzentration der IgE-Antikörper im Blut, auch Titer genannt, als Reaktion auf den Kontakt mit einem Allergen kann dabei als quantitativer Phänotyp aufgefasst werden.

Aus diesem Grund wurde GENEHUNTER-QMOD an einem Datensatz zu einer genomweiten Kandidatenregion-Suche bei der Allergisierung gegen Extrakt von Hausstaubmilben (*Dermatophagoides pteronyssinus*) angewendet. Der Datensatz besteht aus Familien verschiedener europäischer Populationen. Hier wurden die englischen und deutschen Familien untersucht. Es handelt sich dabei hauptsächlich um sib-pair-Familien (die beiden Eltern sowie zwei betroffene Geschwister), aber auch einige komplexere Familien sind enthalten.

Die deutsche Population setzt sich aus 44 Familien mit 187 Individuen zusammen, darunter 33 Familien mit 2 Geschwistern und 11 Familien mit 3 Geschwistern. Die englische Population besteht aus 19 Familien mit 122 Individuen. Hierunter sind

7 Familien mit 2 Geschwistern, 3 Familien mit 3 Geschwistern, 5 Familien mit 4 Geschwistern und 4 Familien mit komplexeren Stammbaumstrukturen.

Für jeden Nonfounder im Datensatz wurde ein sog. RAST-Test (kurz für *radio allergo sorbent test*) durchgeführt. Bei diesem Test wird die Konzentration spezifischer IgE-Antikörper als Reaktion auf hochkonzentrierten Extrakt von *Dermatophagoides pteronyssinus* gemessen. In Abhängigkeit des IgE-Wertes wurde dem Individuum eine RAST-Klasse zugewiesen. Diese reicht von 0 - 6 und bestimmt sich nach dem Schema in Tab. 6.6.

Tabelle 6.6: RAST-Klassen-Zuweisung anhand des IgE-Wertes (Quelle: Wikipedia)

RAST-Klasse	IgE-Wert ($\frac{KU}{L}$)	Kommentar
0	$<$ 0,35	kein oder nicht messbares allergen-spezifisches IgE
1	0,35 - 0,69	geringer allergen-spezifischer IgE-Wert
2	0,70 - 3,49	moderater allergen-spezifischer IgE-Wert
3	3,50 - 17,49	hoher allergen-spezifischer IgE-Wert
4	17,50 - 49,99	sehr hoher allergen-spezifischer IgE-Wert
5	50,00 - 100,00	sehr hoher allergen-spezifischer IgE-Wert
6	$>$ 100,00	extrem hoher allergen-spezifischer IgE-Wert

Die Markerinformationen bestehen aus Genotypen für 604 Mikrosatelliten-Marker, die auf den Chromosomen 1-23 typisiert wurden. Die Resultate der Originalanalysen für den dichotomisierten Hausstaubmilben-Phänotyp sind in Kurz *et al.* (2000) und Kurz *et al.* (2005) beschrieben.

Es gibt Hinweise darauf, dass Parent-of-origin- bzw. Imprinting-Effekte bei der Entwicklung von Allergien eine Rolle spielen (Moffat & Cookson, 1998). Bei der Auswertung des Datensatzes mit GENEHUNTER-QMOD wurde daher das Krankheitsmodell gewählt, welches Imprinting berücksichtigt.

Die MOD-Score-Berechnung mit dem PGRAD-Verfahren wurde mit 10 verschiedenen Startwerten mit anschließender Optimierung durchgeführt. Je ein Startwert wurde nach Methode 1, 2 und 3 aus Kap. 3.3 berechnet, die restlichen sieben Startwerte wurden mittels des Startwert-Zufallsgenerators (Methode 4, Kap. 3.3) erzeugt. Das Ergebnis der MOD-Score-Berechnung war das Maximum der 10 PGRAD-Optimierungen. Zur p-Wert-Berechnung wurde das Stichprobenverfahren benutzt. Die Anzahl der dafür von GENEHUNTER-QMOD erzeugten Replikate unter der Nullhypothese richtete sich nach der Größe der Familien in den einzelnen Populationen,

da dies die Rechenzeit entscheidend beeinflusst. Alle Berechnungen wurden auf dem Marburger Rechencluster (MaRC) durchgeführt.

Die MOD-Score-Analyse mit GENEHUNTER-QMOD lieferte ein positives Resultat auf Chromosom 4, bei Marker D4S194 für die deutsche Population. Der MOD-Score beträgt 1,019, mit einem zugehörigen p-Wert von 0,00124. Dieser p-Wert wurde mit 50.000 Replikaten unter der Nullhypothese mittels des Stichprobenverfahrens berechnet. Ein weiteres Kopplungssignal liegt auf Chromosom 18, bei Marker D18S452, ebenfalls für die deutsche Population. Der MOD-Score ist hier 0,750, der entsprechende p-Wert beträgt 0,043. Für diesen p-Wert wurden 1.000 Nullhypothesenreplikate mit dem Stichprobenverfahren berechnet. Für die englische Population trat ein MOD-Score von 1,193 auf Chromosom 5, bei Marker D5S486, auf. Der zugehörige p-Wert beträgt 0,001. Er wurde mit 1.000 Nullhypothesenreplikaten berechnet.

Alle Genorte wurden bereits in Strauch *et al.* (2000) identifiziert und somit durch die Analyse mit GENEHUNTER-QMOD bestätigt. In Strauch *et al.* (2000) wurde das Programm GENEHUNTER-IMPRINTING verwandt, welches dichotome Phänotypen verwendet. Der quantitative Phänotyp RAST-Klasse wurde entsprechend dichotomisiert. Eine Person mit einer RAST-Klasse von 1 oder höher galt dabei als betroffen, eine Person mit einer RAST-Klasse von 0 entsprechend als nicht betroffen. Die höchsten MOD-Scores lagen hier für die deutsche Population bei den Markern D18S452 und D4S430. Letzterer liegt direkt neben D4S194. Bei der englischen Population lag der MOD-Score bei Marker D5S416, welcher direkt neben D5S486 liegt. Die Ergebnisse finden sich zusammengefasst in Tabelle 6.7.

Tabelle 6.7: Ergebnisse der Kopplungsanalyse der Hausstaubmilben-Daten

Population	Chromosom	Marker	genetische Position	MOD-Score	p-Wert
Deutsch	4	D4S194	121,4 cM	1,019	0,00124
Deutsch	18	D18S452	9,7 cM	0,750	0,043
Englisch	5	D5S486	27,84 cM	1,193	0,001

7 Diskussion

Das Ergebnis des Vergleichs von PGRAD- und Stichprobenverfahren (Szenario 1 und 2, Tabelle 6.1) scheint zu überraschen, da das Szenario mit den kleineren Störtermen die geringere Power hat. Erwarten würde man, dass ein Phänotyp mit größeren familien- und personenspezifischen Restvarianzen kleinere MOD-Scores liefert, da die Kopplung schwieriger zu entdecken ist. Dies wiederum führt dazu, dass sich die MOD-Scores der unter H_1 simulierten Datensätze weniger stark von den H_0-Replikaten unterscheiden. Das hat zur Folge, dass bei einem Vergleich des MOD-Scores für einen unter H_1 simulierten Datensatz mit jenem der daraus erzeugten H_0-Replikate letzterer den der H_1-Datensätze häufiger übersteigt, was zu größeren p-Werten und somit kleinerer Power führt. Die entgegengesetzte Beobachtung erscheint umso erstaunlicher, da Szenario 1 mit dem PGRAD-Verfahren berechnet wurde, einer komplexen mathematischen Optimierungsroutine, während bei Szenario 2 das vergleichsweise simple Stichprobenverfahren benutzt wurde, bei dem es sich letztendlich nur um den Mittelwert einiger Funktionswerte handelt.

Wie also ist die kleinere Power zu erklären? Eine Vermutung lässt sich aus der unterschiedlichen Struktur der MOD-Score-Funktionen von Datensätzen ableiten, die unter H_0 bzw. H_1 simuliert wurden. Die MOD-Score-Funktion der H_1-Datensätze, in denen tatsächlich Kopplung vorliegt, besitzt prinzipiell höhere Funktionswerte als ihr H_0-Gegenstück. Allerdings ist bei einem derartig großen Parameterraum mit 7 bzw. 9 Dimensionen auch die Anzahl der lokalen Maxima hoch. Es gibt viele verschiedene Parameterkombinationen, die die Daten einigermaßen plausibel erklären. Es besteht also die Möglichkeit, dass das PGRAD-Verfahren bei der Maximierung in einem der lokalen Maxima stecken bleibt. Diese Möglichkeit gibt es bei den H_0-Replikaten auch, allerdings ist die Wahrscheinlichkeit dafür kleiner. Da in H_0-Datensätzen keine Kopplung vorliegt, die meisten Parameterkombinationen die Daten also nicht gut erklären können, ist die Funktion überwiegend 0 oder sehr nahe bei 0 und besitzt darum weniger hohe lokale Maxima. Die Wahrscheinlichkeit, mit dem PGRAD-Verfahren das

absolute Maximum zu finden, ist höher. Dieses wiederum kann dann größer sein als ein lokales Maximum des zugehörigen H_1-Datensatzes. Weiterhin werden sich lokale und globale Maxima im H_0-Datensatz nicht so stark voneinander unterscheiden wie im H_1-Datensatz unter Kopplung. Das PGRAD-Verfahren erreicht den "optimalen" MOD-Score unter H_0 also leichter als unter H_1. Dieses Phänomen treibt offenbar unter H_1, d.h. bei Szenarien mit Kopplung, den p-Wert künstlich nach oben und verkleinert die Power.

Das Stichprobenverfahren leidet nicht unter diesem Nachteil. Jeder Parameterbereich der Funktion wird mit gleicher Wahrscheinlichkeit in die Mittelwertberechnung einbezogen. Bei durchschnittlich höheren Funktionswerten in den H_1-Datensätzen schlägt sich das in einem höheren Mittelwert nieder, ungeachtet der Anzahl der Nebenmaxima. Dieser Vorteil und die Tatsache, dass die Berechnung des Mittelwerts um ein Vielfaches schneller erfolgt als die Optimierung durch das PGRAD-Verfahren, machen das Stichprobenverfahren deutlich überlegen, wenn es um die p-Wert-Berechnung eines Datensatzes geht.

Ansonsten ist von der Anwendung der Stichproben-Methode abzuraten. Weder berechnet sie den maximalen MOD-Score, noch liefert sie den Parametervektor w, der zu dem maximalen MOD-Score gehört. Im Hinblick auf die Genotyp-Phänotyp-Relation liefert sie also keine Erkenntnis. Hier ist stattdessen das PGRAD-Verfahren zu benutzen.

Powervergleiche zwischen dem Stichprobenverfahren und der Varianzkomponentenanalyse für Szenario 3 und 5 zeigen, dass die VCA besser abschneidet, wenn normalverteilte Phänotypen und kleine Stammbäume analysiert werden. In Szenario 3 (gleiche Restvarianzen für alle drei Genotypen) beträgt die Power des Stichprobenverfahrens 82%, die der VCA 98%; letztere ist damit um 16% besser. Man beachte, dass die VCA ebenfalls gleiche Varianzen voraussetzt und es nicht möglich ist, ein Modell mit verschiedenen Varianzen zu wählen. Auch die Haseman-Elston-Regression ist hier mit einer Power von 91,5% dem Stichprobenverfahren überlegen.

In Szenario 5 wurden verschiedene Varianzen gewählt; es wurde also ein Modell benutzt, dass dem Stichprobenverfahren zupaß kommt. Dementsprechend sinkt die Power der VCA ein wenig auf 95%, das Stichprobenverfahren verbessert sich auf 87,4%. Die Power der VCA ist also ebenfalls höher, die Differenz beträgt aber nur noch 7,6%. Die Haseman-Elston-Regression liefert hier mit 74,5% die schlechteste Power. Eine Möglichkeit zur Erklärung der Ergebnisse liegt im Modell und

der Optimierung des Stichprobenverfahrens: Die Genotyp-Phänotyp-Relation ist bei der Modellierung genotypspezifischer Restvarianzen sehr allgemein formuliert und deswegen recht komplex. Die hohe Anzahl an Parametern, die das Modell enthält $\big(\mu_{(+,+)}, \mu_{(m,+)}, \mu_{(+,+)}, \sigma_{(+,+)}, \sigma_{(m,+)}, \sigma_{(+,+)}\big)$, hat das Potential, verschiedenste Phänotypkonstellationen zu erklären. Allerdings erwächst daraus eine komplexere LOD-Score-Funktion als bei gleichen Restvarianzen. Das Maximum dieser Funktion zu finden, gestaltet sich schwieriger als bei Funktionen mit weniger Parametern. Die Schwierigkeit für die Haseman-Elston-Regression besteht in diesem Szenario darin, dass eine zunehmende genotypische Ähnlichkeit (IBD-Status) nicht zwangsläufig eine phänotypische Ähnlichkeit zur Folge hat. Bei einem IBD-Status von 2, der hier einen identischen Genotyp zur Folge hat (Marker liegt direkt auf dem Krankheitslocus), können sich die Phänotypen je nach Genotyp immer noch stark unterscheiden, da die Varianz des (m, m)-Genotyps hoch gewählt wurde. Dies mindert die Korrelation zwischen IBD-Status und quadrierter phänotypischer Differenz.

Analysiert man Daten mit größeren Stammbäumen (Szenario 4), so schneidet das Stichprobenverfahren besser ab als die VCA und die Haseman-Elston-Regression. In unserem Beispiel war das Stichprobenverfahren in der Lage, den gegenüber Szenario 3 stärkeren Störterm mit der Information der zusätzlichen Personen im Stammbaum zu kompensieren und weiterhin eine Power von 82% zu liefern. Die Varianzkomponentenanalyse und die Haseman-Elston-Regression fielen hier in der Power deutlich zurück. Insbesondere die Haseman-Elston-Regression hat mit 22,5% eine deutlich zu niedrige Power. Es scheint, als wären VCA und die Haseman-Elston-Regression nicht so gut in der Lage, die zusätzlichen Informationen in erweiterten Stammbäumen auszunutzen wie das Stichprobenverfahren, oder kommen mit stärkeren Störtermen schlechter zurecht. Bei der Haseman-Elston-Regression ist das nicht überraschend, da größere Stammbäume in alle mögliche Geschwisterpaarkombinationen zerlegt werden.

Ein weiterer Vorteil des Stichprobenverfahrens zeigt sich, wenn ein nicht normalverteilter Phänotyp analysiert wird. Sowohl das Stichprobenverfahren als auch die Varianzkomponentenanalyse setzen eine genotypspezifische Normalverteilung des Phänotyps voraus. Diese Voraussetzung ist offensichtlich nicht erfüllt. Eigentlich wäre hier bei beiden Verfahren ein Abfall der Power gegenüber dem Szenario mit Normalverteilung zu erwarten. Es zeigt sich jedoch, dass das Stichprobenverfahren besser mit nicht normalverteilten Phänotypen zurechtkommt als die VCA und eine um 35%

höhere Power liefert. Es scheint, als ob durch die Modellierung einer eigenen Verteilung für jeden Genotyp und die dadurch hohe Anzahl an "Erklärungsmöglichkeiten" für den Phänotyp eine gewisse Robustheit im Modell gegenüber einer Verletzung der Normalverteilungsannahme vorliegt. Ist man in der realen Anwendung also unsicher, welcher Verteilung die Phänotypen folgen, bietet es sich an, das Stichproben/PGRAD-Verfahren anstelle der VCA zu benutzen. Obwohl die Haseman-Elston-Regression modell-frei ist, also keine Normalverteilung voraussetzt, kommt sie mit den Daten aus Szenario 6 noch schlechter zurecht und liefert mit 13% die geringste Power, was den Schluss nahe legt, dass die Haseman-Elston-Regression für diese Art von Daten ungeeignet ist.

Auch bei der Betrachtung des Fehlers 1. Art ist die VCA bei normalverteilten Daten (Szenario 3 und 5) etwas präziser, d.h. der tatsächliche Fehler 1. Art ist etwas näher am theoretischen Wert von 5%. Szenario 5 zeigt, dass das Stichprobenverfahren manchmal zu konservativ ist, also zu selten die Nullhypothese verwirft. Das kann bei Simulationen unter H_1 zu einer reduzierten Power führen. Möglicherweise ist dies ein Grund dafür, warum die Power des Verfahrens gegenüber der VCA bei kleinen Stammbäumen und normalverteilten Phänotypen bei zufälliger Rekrutierung leicht zurückliegt.

Szenario 6 zeigt, dass auch unter der Nullhypothese keiner Kopplung das Stichprobenverfahren mit nicht normalverteilten Daten besser umgehen kann als die VCA. Das Stichprobenverfahren verwirft hier mit 3,6% zwar zu selten, hält jedoch das nominale Testniveau von 5% ein. Die VCA hat einen tatsächlichen Fehler 1. Art von 13,4% und ist damit deutlich zu liberal. Die Haseman-Elston-Regression hält hier das Niveau durchweg gut ein. Allerdings wird sie bei "komplizierteren" Szenarien (verschiedene genotypspezifische Restvarianzen, keine Normalverteilung) etwas konservativ, was sich entsprechend unter H_1 in einer kleineren Power niederschlägt.

In Datensätzen mit nicht-zufälligem Rekrutierungsschema zeigt das Stichprobenverfahren eine überlegene Power in allen drei Szenarien. Der Varianzkomponentenanalyse misslingt vollständig die Analyse des *single-proband-selection*-Szenarios und des *double-proband-selection*-Szenarios mit Phänotypen der Geschwister in entgegengesetzten Quartilen. Lediglich bei *double-proband-selection* mit Geschwistern in gleichen Quartilen ist die Power akzeptabel, was jedoch nur durch einen überhöhten tatsächlichen Fehler 1. Art erreicht wird. Vom Gebrauch der VCA sollte also bei solchen Daten Abstand genommen werden. Die Haseman-Elston-Regression bleibt hinsicht-

lich der Power in allen Szenarien deutlich hinter der Power des Stichprobenverfahrens zurück, obwohl das Verfahren einen akkuraten tatsächlichen Fehler 1. Art liefert. Man sollte das Stichprobenverfahren hier also auch der Haseman-Elston-Regression vorziehen.

Bezüglich des Fehlers 1. Art und der Power decken sich die Resultate zu VCA und zur Haseman-Elston-Regression in dieser Arbeit zu großen Teilen mit Kleensang *et al.* (2010). Auch dort weicht die VCA oft vom nominalen Testniveau ab und ist deutlich zu liberal, wenn die Bedingung eines normalverteilten Phänotyps verletzt ist. Der Effekt einer Ascertainment-Bedingung beeinflusst das Testniveau ebenfalls stark, und zwar je nach Rekrutierungsschema in beide Richtungen. Es wird ebenfalls eine geringe Power bei nicht normalverteilten Phänotypen beschrieben. Die Haseman-Elston-Regression ist konservativ bei nicht normalverteilten Phänotypen, was die hier erhaltenen Resultate bestätigen. Die Power wird als der Power der VCA unterlegen beschrieben, wenn Rekrutierungsbedingungen gelten. Dies ist ebenfalls konsistent mit unseren Ergebnissen.

Die Parameterschätzung zeigt relativ gute Resultate für die Erwartungswerte μ_i der Normalverteilungen f_i für die einzelnen Genotypen. In Szenario 3 ist insbesondere der Schätzer des Mittelwertes für den homozygot mutanten Genotyp $\mu_{(m,m)}$ mit nur 2% relativer Abweichung, bezogen auf die Differenz der Mittelwerte, präzise. Die relative Abweichung von 13% des Schätzers für $\mu_{(m,+)}$ ist auch noch akzeptabel. Mit einer Abweichung von 24,5% wird $\mu_{(+,+)}$ allerdings merklich unterschätzt. Der Schätzer für die Standardabweichungen ist deutlich zu niedrig. Der Fehler für die einzelnen Schätzer änderte sich nicht merklich, wenn ein nicht zufälliges Rekrutierungsschema vorlag (Szenario 8). Auch hier ist die Abweichung des Schätzers für $\mu_{(m,m)}$ am kleinsten, und der Schätzer für die Standardabweichungen ist deutlich zu niedrig.

Die Histogramme der Erwartungswerte $\mu_{(+,+)}$, $\mu_{(m,+)}$ und $\mu_{(m,m)}$ für Szenario 3 (Bild 6.2, 6.3 und 6.4) zeigen ein deutliches Maximum in der Nähe des jeweils wahren Wertes. Zu beiden Seiten dieses Maximums flachen die Balken deutlich ab. Bei der Anwendung auf reale Daten zeigt sich, dass die MOD-Score-Analyse mit GENEHUNTER-QMOD auch in der Praxis potentiell krankheitsverursachende Genorte identifizieren kann. Dies wird dadurch verifiziert, dass eine vorhergehende Kopplungsanalyse mit einer dichotomen Phänotypdefinition Kopplungssignale an denselben oder direkt benachbarten Markern wie die Analyse mit GENEHUNTER-QMOD ergab.

8 Ausblick

Die Kopplungsanalyse ist eine wichtige Methode, um Gene zu identifizieren, die Erbkrankheiten hervorrufen oder ursächlich an ihrer Entstehung beteiligt sind. Dies trifft auch im Zeitalter der Kartierung seltener genetischer Varianten mittels Hochdurchsatz-Sequenzierung zu. Hier hat die Kopplungsanalyse die wichtige Aufgabe, genetische Regionen, in denen genetische Unterschiede zwischen Betroffenen und Gesunden durch Sequenzierung identifiziert wurden, weiter einzugrenzen. Indem die entsprechende genetische Variante und ihr Einfluss auf die Krankheit ermittelt werden, kann der Krankheitsmechanismus aufgeklärt werden, was die Möglichkeit eröffnet, die Ursache der Krankheit zu behandeln und nicht nur deren Symptome. Ebenfalls ist es möglich, eine Krankheit bereits vor ihrem Ausbruch zu diagnostizieren. Auf diese Weise können rechtzeitig vorbeugende Maßnahmen ergriffen werden.

Mit GENEHUNTER-QMOD wird dem Spektrum an Kopplungsanalyse-Methoden ein Verfahren für die Analyse quantitativer Phänotypen hinzugefügt. Das Verfahren führt sowohl einen statistischen Signifikanztest auf Kopplung als auch eine Schätzung der Parameter der Genotyp-Phänotyp-Relation durch. Der Phänotyp wird dabei als eine genotypspezifische, normalverteilte Zufallsvariable modelliert. Die Schätzer beinhalten sowohl die Mittelwerte als auch die Standardabweichungen der genotypspezifischen Normalverteilungen. Wenn weitere Informationen über den Vererbungsmodus der Krankheit zur Verfügung stehen, kann das Krankheitsmodell entsprechend angepasst werden: Es ist möglich, Imprinting, keine Dominanzeffekte und gleiche Restvarianzen für alle Genotypen bei der Kopplungsanalyse zu berücksichtigen.

Die Berechnung der Teststatistik, des MOD-Scores, führt zu einem komplexen numerischen Problem: Zum einen ist die Berechnung des LOD-Scores in GENEHUNTER bei größeren Stammbäumen rechenaufwendig. Zum anderen muss die LOD-Score-Funktion über bis zu neun Parameter maximiert werden, das heißt, ein hochdimensionales Optimierungsproblem ist zu lösen. Dies wird zusätzlich dadurch erschwert, dass die Maximierung auf beschränkten Teilräumen des $\mathbb{R}^n$ erfolgen muss, da sich für

die Optimierungsparameter natürliche obere und untere Schranken ergeben. Weiterhin steigt durch empirische Berechnung des p-Wertes, der einem bestimmten MOD-Score entspricht, der absolute Rechenaufwand einer Kopplungsanalyse um Größenordnungen von bis zu 10^5. Ein solches Problem stellt hohe Ansprüche an mehrdimensionale Optimierungsverfahren. Es hat sich jedoch gezeigt, dass eine Kombination aus PGRAD- und Stichprobenverfahren diesen Ansprüchen gerecht werden kann. Das Optimierungsproblem kann durch geschickte Wahl und Anzahl der Startwerte mit dem PGRAD-Verfahren in akzeptabler Zeit gelöst werden. Dem Rechenzeitbedarf der p-Wert-Berechnung kann mit dem Stichprobenverfahren begegnet werden. Dieses führt zwar keine Maximierung des LOD-Scores durch und gestattet keine Parameterschätzung, liefert jedoch die relevanten Informationen für den p-Wert und ist um ein Vielfaches schneller als die PGRAD-Optimierung. Unter Umständen erhält man damit sogar die höhere Power.

Das Verfahren eignet sich besonders gut, wenn die Annahme eines normalverteilten Phänotyps nicht mit Sicherheit getroffen werden kann, oder wenn nicht zufällig rekrutierte Daten analysiert werden. Auch bei der Analyse komplexerer Stammbäume ist das Verfahren anderen Methoden überlegen, auch wenn die Rechenzeit deutlich erhöht ist. Weiterhin liefert das PGRAD-Verfahren durch die Parameterschätzung umfassendere Informationen über die Genotyp-Phänotyp-Relation der Erbkrankheit als andere Verfahren.

Neben der Identifikation des Gens per se können die erhaltenen Informationen auf mehrere Arten nützlich sein. Wenn sich der durchschnittliche Phänotyp deutlich mit dem Genotyp ändert, könnten Patienten mit unterschiedlichen Genotypen individualisierte Behandlungen benötigen, wie z.B. eine andere Medikation oder eine unterschiedliche Medikamentendosis. Die Schätzer der Krankheitsmodellparameter können weitere Hinweise auf den Krankheitsmechanismus geben. Falls sich die Restvarianz mit dem Genotyp ändert, könnten speziell bei Vorliegen eines oder zweier Mutationsallele am Krankheitslocus weitere Mechanismen zum Tragen kommen, die den Phänotyp beeinflussen, z.B. andere Gene oder Umweltfaktoren, die nun einen stärkeren Einfluss haben.

Da GENEHUNTER-QMOD auf dem Lander-Green-Algorithmus basiert, können viele Marker gleichzeitig in die Kopplungsanalyse einbezogen werden. Darum eignet sich GENEHUNTER-QMOD gut für die Anwendung in Genkartierungsprojekten mit diallelischen SNP-Markern, die weniger informativ sind als Mikrosatelliten und daher in

größerer Zahl in die Analyse eingehen müssen.

Die Speicher- und Rechenkapazitäten zukünftiger Computergenerationen könnten es ermöglichen, die Leistung des PGRAD-Verfahrens weiter zu steigern. So liefert eine Erhöhung der maximalen Iterationszahl oder der Anzahl der zu testenden Startwerte eine gründlichere Optimierung des MOD-Scores, was zu präziseren Parameterschätzungen und höherer Power führen kann.

GENEHUNTER-QMOD ist nicht kommerzielle Software und im Internet unter *http://www.helmholtz-muenchen.de/genepi/downloads* frei erhältlich.

Zusammenfassung

Motivation: Krankheiten beim Menschen werden zu einem großen Teil durch genetische Varianten beeinflusst oder verursacht. Um den Krankheitsmechanismus zu verstehen und um Patienten ursächlich behandeln zu können, ist ein erster Schritt, die genetische Variante im menschlichen Genom zu lokalisieren. Ein wichtiges Hilfsmittel für dieses Ziel ist die *Kopplungsanalyse*. In dieser Arbeit wird ein neues parametrisches Verfahren für die Analyse quantitativer Phänotypen vorgestellt. Das Verfahren führt sowohl einen statistischen Signifikanztest als auch eine Schätzung verschiedener Parameter der Genotyp-Phänotyp-Relation durch. Das Verfahren wurde in das Software-Paket GENEHUNTER-QMOD implementiert und anhand von simulierten Daten mit bestehenden Verfahren, nämlich der Varianzkomponentenanalyse (VCA) und der Haseman-Elston-Regression, bzgl. Power und Fehlerniveau 1. Art verglichen.
Methoden: Der Phänotyp wird als eine normalverteilte Zufallsvariable modelliert, mit einer eigenen Verteilung für jeden Genotyp. Die Schätzer der genotypspezifischen Erwartungswerte und Standardabweichungen erhält man, indem man den LOD-Score mit einer gradientenbasierten Optimierungsmethode, dem PGRAD-Verfahren, über diese Parameter maximiert. Auf diese Weise kann GENEHUNTER-QMOD sowohl den putativen Krankheitslocus ermitteln als auch spezifische Informationen über die Genotyp-Phänotyp-Relation liefern.
Ergebnisse: Im Falle einer Normalverteilung und bei Geschwisterpaaren hat das PGRAD-Verfahren eine geringere Power als die VCA. Allerdings ist es bei größeren Stammbäumen, nicht zufällig rekrutierten Daten oder nicht normalverteilten Phänotypen sowohl der VCA als auch der Haseman-Elston-Regression überlegen. Im letzteren Fall geht die höhere Power sogar mit einem reduzierten Fehlerniveau 1. Art einher, während die VCA zu liberal ist. Bei der Bestimmung der Genotyp-Phänotyp-Relation werden die Restvarianzen unterschätzt, die Schätzer der Erwartungswerte der Phänotyp-Verteilungen sind präziser.
Schlussfolgerung: Mit GENEHUNTER-QMOD wird das Spektrum der Kopplungsanalyse-

Methoden um ein mächtiges Verfahren zur genetischen Kartierung quantitativer Phänotypen erweitert. Da GENEHUNTER-QMOD auf dem Lander-Green-Algorithmus basiert, können viele Marker gleichzeitig in die Kopplungsanalyse einbezogen werden. Darum eignet sich GENEHUNTER-QMOD gut für die Anwendung in Genkartierungsprojekten mit diallelischen SNP-Markern, die weniger informativ sind als Mikrosatelliten und daher in größerer Zahl in die Analyse eingehen müssen. GENEHUNTER-QMOD ist nicht kommerzielle Software und im Internet unter *http://www.helmholtz-muenchen.de/genepi/downloads* frei erhältlich.

Summary

Objective: To a large degree human diseases are influenced or caused by genetic variants. In order to understand the mechanism of the disease and to treat patients in a causative way, a first step is to locate the genetic variants in the human genome. An important tool for this goal is *linkage analysis*. In this work, a parametric method for linkage analysis of quantitative phenotypes is presented. The method provides a test for linkage as well as an estimate of different parameters of the genotype-phenotype relation. We have implemented our new method in the program GENEHUNTER-QMOD and performed simulations to compare its power and type I error to existing methods, i.e. variance components analysis (VCA) and Haseman-Elston regression.

Methods: The phenotype is modeled as a normally distributed variable, with a separate distribution for each genotype. Estimates of the genotype-specific expectation values and standard deviations are obtained by maximizing the LOD score over these parameters with a gradient-based optimization called PGRAD method. That way, GENEHUNTER-QMOD can both locate the putative disease locus and provide specific information about the genotype-phenotype relation.

Results: GENEHUNTER-QMOD has lower power to detect linkage than VCA in case of a normal distribution and with sib pairs. However, it outperforms VCA and Haseman-Elston regression for larger pedigrees, non-randomly ascertained data or non-normally distributed phenotypes. Here, the higher power even goes along with conservativeness, while VCA has an inflated type I error. Parameter estimation tends to underestimate residual variances, but performs better for expectation values of the genotype-specific phenotype distributions.

Conclusion: With GENEHUNTER-QMOD, a powerful new tool is provided to explicitly model quantitative phenotypes in the context of linkage analysis. Because GENEHUNTER-QMOD is based on the Lander-Green algorithm, it can simultaneously use many markers in the analysis, which makes it applicable to gene-mapping projects based on SNP arrays. The program is freely available at
http://www.helmholtz-muenchen.de/genepi/downloads.

Literaturverzeichnis

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

Almasy, L. & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet*, 62:1198–1211.

Birgin, E. & Martínez, J. (2002). Large-scale active-set box-constrained optimization method with spectral projected gradients. *Computational Optimization and Applications*, 23:101–125.

Birgin, E., Martínez, J. & Raydan, M. (2000). Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10:1196–1211.

Camp, N. J. & Cox, A. (2002). *Quantitative Trait Loci*. humanapress, Totowa, New Jersey.

Cottingham, R., Idury, R. & Schäffer, A. (1993). Faster sequential genetic linkage computations. *Am J Hum Genet*, 53(1):252–263.

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

Dietter, J., Spiegel, A., an Mey, D., Pflug, H.-J., Al-Kateb, H., Hoffmann, K., Wienker, T. & Strauch, K. (2004). Efficient two-trait-locus linkage analysis through program optimization and parallelization: application to hypercholesterolemia. *Eur J Hum Genet*, 12:542–550.

Dudbridge, F. (2003). A survey of current software for linkage analysis. *Human Genomics*, 1(1):63–65.

Elston, R. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Hum Hered*, 21:523–542.

Falconer, D. S. & Mackay, F. C. (1996). *Introduction to Quantitative Genetics*. Pearson Education, Harlow, England.

Franke, D., Kleensang, A., Elston, R. C. & Ziegler, A. (2005). Haseman-elston weighted by marker informativity. In *Genetic Analysis Workshop 14: Microsatellite and single-nucleotide polymorphism*.

Franke, D., Kleesang, A. & Ziegler, A. (2006). Sibsim - quantitative phenotype simulation in extended pedigrees. *GMS Med Inform Biom Epidemiol.*

Gudbjartsson, D., Jonasson, K., Frigge, M. & Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nat Genet*, 25:12–13.

Haseman, J. K. & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1).

Hasstedt, S. & Cartwright, P. (1981). Pap-pedigree analysis package. technical report no. 13. *Technical report, Department of Medical Biophysics and Computing, University of Utah*.

Heuser, H. (2004). *Lehrbuch der Analysis Teil 1.* Teubner, Wiesbaden.

Heuser, H. (2006). *Lehrbuch der Analysis Teil 2.* Teubner, Wiesbaden.

Kelley, C. T. (1999). *Iterative Methods for Optimization*. SIAM Society for Industrial and Applied Mathematics, Philadelphia.

Kleensang, A., Franke, D., Alcaïs, A., Abel, L., Müller-Myshok, B. & Ziegler, A. (2010). An extensive comparison of quantitative trait loci mapping methods. *Hum Hered*, 69:202–211.

Kong, A. & Cox, N. (1997). Allele-sharing models: Lod scores and accurate linkage tests. *Am J Hum Genet*, 61:1179–1188.

Kruglyak, L., Daly, M., Reeve-Daly, M. & Lander, E. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet*, 58:1347–13636.

Kurz, T., Altmueller, J., Strauch, K., Rüschendorf, F., Heinzmann, A., Moffat, M., Cookson, W., Inacio, F., Nürnberg, P., Stassen, H. & Deichmann, K. (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., Strauch, K., Heinzmann, A., Braun, S., Jung, M., Rüschendorf, F., Mofatt, M., Cookson, W., Inacio, F., Ruffilli, A., Nordskov-Hansen, G., Peltre, G., Forster, J., Kuehr, J., Reis, A., Wienker, T. & Deichmann, K. (2000). A european study on the genetics of mite sensitization. *J Allergy Clin Immunol*, 106:925–932.

Lander, E. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci*, 84:2363–2367.

Lathrop, G. & Lalouel, J. (1984). Easy calculations of lod scores and genetic risks on small computers. *Am J Hum Genet*, 36(2):460–465.

Mattheisen, M., Dietter, J., Knapp, M., Baur, M. & Strauch, K. (2008). Inferential testing for linkage with GENEHUNTER-MODSCORE: The impact of the pedigree structure on the null distribution of multipoint mod scores. *Genetic Epidemiology*, 32:73–83.

Moffat, M. & Cookson, W. (1998). Maternal effects in atopic disease. *Clin Exp Allergy*, 28:56–61.

Pratt, C., Daly, M. J. & Kruglyak, L. (2000). Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. *Am J Hum Genet*, 66(1):1153–1157.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rowe, S. J. (2008). *QTL mapping technology using variance components in general pedigrees applied to the poulty industry*. Diplomarbeit, University of Edinburgh.

SIAM (1996). Numerical methods for unconstrained optimization and nonlinear equations. In *Classics in Applied Mathematics*, Band 16. SIAM, Philadelphia.

Strauch, K. (2002). *Kopplungsanalyse bei genetisch komplexen Erkrankungen mit genomischem Imprinting und Zwei-Genort-Krankheitsmodellen*. Urban und Vogel (Medizinische Informatik, Biometrie und Epidemiologie, Medizin & Wissen, Bd. 87), München.

Strauch, K. (2007). Mod-score analysis with simple pedigrees: An overview of likelihood-based linkage methods. *Hum Hered*, 64:192–202.

Strauch, K., Fimmers, R., Kurz, T., Deichmann, K., Wienker, T. & Baur, M. (2000). Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: Application to mite sensitization. *Am J Hum Genet*, 66:1945–1957.

Zhao, H., Merikangas, K. & Kidd, K. (1999). On a randomization procedure in linkage analysis. *Am J Hum Genet*, 65:1449–1456.

Ziegler, A. (1999). *Genetische Kartierung quantitativer Phänotypen*. Urban und Vogel (Medizinische Informatik, Biometrie und Epidemiologie, Medizin & Wissen, Bd. 84), München.

Ziegler, A. & Kastner, C. (1997). A minimum distance estimation approach to estimate the recombination fraction from a marker locus in robust linkage analysis for quantitative traits. *Biometrical Journal*, 39:765–775.

Danksagung

Mein persönlicher Dank gebührt vorrangig Prof. Konstantin Strauch für die hervorragende Betreuung meiner Doktorarbeit, insbesondere gegen Ende der Arbeit trotz großer räumlicher Distanz. Weiterhin möchte ich mich bei dem gesamten Institut für Medizinische Biometrie und Epidemiologie in Marburg für das angenehme Arbeitsklima und die kollegiale Hilfe bedanken. Nicht zuletzt bei Prof. Schäfer für die freundliche Aufnahme, als meine Arbeitsgruppe nach München abgewandert ist...

Zusätzlich möchte ich mich herzlich bei Dr. Thorsten Kurz und Prof. Dr. Klaus Deichmann, vormals Universitäts-Kinderklinik Freiburg, für die Bereitstellung der Daten über Hausstaubmilben-Allergie bedanken.

Die Simulationen mit GENEHUNTER-QMOD wurden auf dem MaRC-Cluster der Philipps-Universität Marburg durchgeführt.

Diese Arbeit ist im Rahmen des DFG-Projektes "Weiterentwicklung und Implementation von Methoden der Kopplungsanalyse für die genetische Kartierung komplexer Krankheiten" (Str643/4-1) entstanden.