Aus dem Institut für Medizinische Biometrie und Statistik
der Universität zu Lübeck
Direktor: Prof. Dr. rer. nat. Andreas Ziegler

---

# Genetische Kartierung quantitativer Merkmale – Ein Gütevergleich kopplungsanalytischer Verfahren

Inauguraldissertation

zur
Erlangung der Doktorwürde
der Universität zu Lübeck

**- Aus der Medizinischen Fakultät -**

vorgelegt von
André Kleensang
aus Hamburg

Lübeck 2010

1. Berichterstatter: Prof. Dr. rer. nat. Andreas Ziegler

2. Berichterstatter: Prof. Dr. med. Gabriele Gillessen-Kaesbach

Tag der mündlichen Prüfung: 05.07.2010

zum Druck genehmigt. Lübeck, den 05.07.2010

gez. Prof. Dr. med. Werner Solbach

- Dekan der Medizinischen Fakultät -

| | |
|---|---|
| ' | Transponiert |
| add | Additiv (kodominant) |
| COAG | Consortium on Asthma Genetics |
| cM | centi Morgan |
| *g* | Hauptgeneffekt |
| *G* | Polygener Effekt |
| dom | Dominant |
| EM | Expectation maximization |
| $\varepsilon$ | Fehlerterm |
| ESP | Doppelte Selektion von Geschwisterpaaren oder extreme Geschwisterpaar-Ansätze |
| Gh.HE.Trad | HASEMAN-ELSTON Verfahren, Implementation in `Genehunter` |
| Gh.VC | VARIANZKOMPONENTENMODELLE, Implementation in `Genehunter` |
| HE | HASEMAN-ELSTON Verfahren |
| IBD | Identical by descent, identisch durch Abstammung |
| IMBS | Institut für Medizinische Biometrie und Statistik der Medizinischen Universität zu Lübeck |
| Linkage | MODELBASIERTES Verfahren mit dem `Linkage` Softwarepaket |
| LDL | Lipoprotein-Fraktion mit geringer Dichte (light density lipoproteine) |
| log NV | Logarithmierte Normalverteilung |
| LQT | Likelihood-Quotienten Test |
| Merlin.K&C | MERLIN-QTL Verfahren, Teststatistik nach Kong und Cox (1997) |
| Merlin.W&H | MERLIN-QTL Verfahren, Teststatistik nach Whittemore und Halpern (1994) |
| MERLIN-REGRESS | Verfahren von Sham et al. (2002) |
| MLB | MAXIMUM LIKELIHOOD BINOMIAL Verfahren |
| MLBQT | MAXIMUM LIKELIHOOD BINOMIAL Verfahren für quantitative Phänotypen |
| MLBQT.Kat | MAXIMUM LIKELIHOOD BINOMIAL Verfahren für quantitative Phänotypen unter Verwendung von empirischen Dezilen der populationsbasierten Verteilung der Phänotypen |
| MLBQT.NV | MAXIMUM LIKELIHOOD BINOMIAL Verfahren für quantitative Phänotypen unter Annahme einer Standardnormalverteilung der Verteilung der Phänotypen |
| Npar | WILCOXON-RANGSUMMENTEST |
| NV | Normalverteilung |
| QTL | Quantitative trait loci, quantitativer Phänotyp Locus |
| rez | Rezessiv |
| rHE | REVIDIERTES HASEMAN-ELSTON Verfahren |

| | |
|---|---|
| RSP | Random sib-pair, zufällig ausgewählte Geschwisterpaare |
| Sage.HE | HASEMAN-ELSTON Verfahren, generalisierte Kleinst-Quadrate-Regression, Implementation in `S.A.G.E.` |
| SPSP | Single proband sib-pair, einfache Selektion von Geschwisterpaaren oder extreme Probanden-Ansätze |
| $\theta$ | Rekombinationsfrequenz |
| VC | VARIANZKOMPONENTENMODELLE |

# 1 Einleitung

## 1.1 Genetische Kartierung quantitativer Merkmale

Quantitative Merkmale, deren Vererbungsmuster auf eine genetische Komponente schließen lassen, ohne einem klaren Mendelschen Erbgang zu folgen, erregen nicht erst in letzter Zeit das besondere Interesse der Humangenetik. Schon zu Beginn des 20. Jahrhunderts diskutierte die wissenschaftliche Gemeinde heftig die scheinbare Unvereinbarkeit der gerade wiederentdeckten Arbeiten Gregor Mendels mit der neu aufkommenden Lehre der Biometrie, deren Gegenstand damals die Vermessung quantitativer Merkmale von Lebewesen und deren statistische Analysen war.

Die Anhänger der Mendelschen Lehre, insbesondere William Bateson, und die insbesondere von den Arbeiten von Sir Francis Galton inspirierten Jünger der Biometrie, vor allem W. F. R. Weldon und später Karl Pearson, diskutierten die damit zusammenhängenden Fragen heftig und teilweise geradezu emotional (Für eine detaillierte Diskussion siehe z.B. Gillham, 2001). Die Biometrie warf der Mendelschen Lehre vor allem vor, sie sei nur sehr begrenzt anwendbar. Die meisten Merkmale seien nämlich quantitativer Art; dies lasse sich aber mit der Mendelschen Lehre nicht vereinbaren.

Schon früh verschafften sich jedoch auch vermitteltende Stimmen Gehör: Bereits 1918 stellte R. A. Fisher eine wissenschaftliche Arbeit vor, die unter anderem die Mendelsche Theorie mit der Biometrie-Lehre in Einklang zu bringen versuchte (Fisher, 1918). Fisher demonstrierte, dass sich quantitative Merkmale und deren erhöhte Korrelationen innerhalb von Familien, wie sie die Biometrie beschreibt, sehr wohl auf der Basis der Mendelschen Lehren erklären lassen, nämlich mit dem Einfluss mehrerer unabhängiger Mendelscher Faktoren. Die Arbeiten von Fisher führte neben anderen Falconer weiter und entwickelte dabei das heute gebräuchliche Modell der polygenen Theorie von quantitativen Merkmalen (Falconer und Mackay, 1996).

Seit Fishers und Falconers grundlegenden Arbeiten wurden mittels genetischer Kartierung quantitativer Phänotypen im Menschen hunderte von chromosomalen Regionen kartiert, die mit einer Vielzahl von Merkmalen oder Erkrankungen wie z.B. Körpergewicht (Rankinen et al., 2006), Körperhöhe (Perola et al., 2007), Knochendichte (Streeten et al., 2006; Zmuda et al., 2006; Perola et al., 2007) oder Malaria (Timmann et al., 2007) zusammenhängen. Erstmals im Jahr 1991 wurde ein Gen erfolgreich unter Zuhilfenahme eines quantitativen Merkmals kartiert und identifiziert (Goate et al., 1991). Inzwischen wurden zahlreiche Gene durch die Verwendung quantitativer Merkmale unmittelbar oder mittelbar identifiziert (Korstanje und Paigen, 2002). Blangero (2004) beschrieb den Erfolg der genetischen Kartierung quantitativer Merkmale im Titel einer Veröffentlichung treffend mit den Worten:„[...] king harvest has surely come.“

Viele Erkrankungen oder Merkmale lassen sich indirekt auf einer quantitativen Skala als intermediäre Phänotypen oder direkt als quantitative Merkmale messen (siehe Tabelle 1). Die quantitativen intermediären Phänotypen ermöglichen im Allgemeinen eine präzisere Definition der Erkrankung bzw. des Merkmals als dichotome Merkmale, die z.B. über einen Referenzbereich definiert werden (Duggirala et al., 1997). Darüber hinaus besitzen statistische Verfahren, die auf quantitativen Merkmalsausprägungen beruhen, im allgemeinen eine höhere statistische Macht (Power) als statistische Verfahren, bei denen die quantitative Größe auf eine dichotome Variable reduziert wird.

**Tabelle 1:** Beispiele für Erkrankungen und quantitative Merkmale als klinisch relevante Phänotypen.

| Erkrankung | Quantitatives Merkmal |
|---|---|
| Artheriosklerose | Cholesterol, Lipoproteine |
| Asthma | IgE |
| Bluthochdruck | Blutdruck |
| Dyslexie | Lese-, Schreibfähigkeit |
| Osteoporose | Knochendichte |
| Übergewicht | Body-mass-index |

Ein besonderes Augenmerk ist bei der genetischen Kartierung quantitativer Merkmale auf das Studiendesign und die Auswahl der kopplungsanalytischen Verfahren zu richten (Terwilliger und Goring, 2000).

Üblicherweise werden zur genetischen Kartierung von quantitativen Merkmalen Kernfamilien, d.h. Eltern und ihre Kinder, rekrutiert. Dem liegt die Überlegung zugrunde, dass eventuell den Phänotyp beeinflussende Umwelteffekte wie Ernährung, Erziehung, allgemeine Lebensumstände sowie weitere mögliche Einflussfaktoren der Umgebung bei den Geschwistern vergleichbar stark sein sollten. Die Familien werden dabei entweder zufällig (RSP, random sib-pair), auf Basis eines phänotypisch extremen Geschwister (SPSP, single proband sib-pair) oder zweier phänotypisch extremen Geschwistern (ESP, extreme sib-pair) ausgewählt. Eine Übersicht der Studiendesigns ist z.B. in Ziegler und König (2010, Kapitel 9) dargestellt.

Bereits im Jahre 1985 stellten Blackwelder und Elston die Vermutung auf, dass die Power statistischer Verfahren zur Kopplungsanalyse erhöht werden könnte, wenn man die Analyse auf Geschwisterpaare begrenzt, bei denen mindestens ein Geschwister eine extreme phänotypische Ausprägung aufweist (Blackwelder und Elston, 1985). Zahlreiche auf diesen Überlegungen basierende Untersuchungen zeigten im wesentlichen, dass bei gleicher Stichprobengröße die statistische Power höher sein kann, wenn über Extremwerte der Phänotypen selektierte Familien untersucht wurden, als bei unselektierten Stichproben (siehe z.B. Carey und Williamson, 1991; Fulker et al., 1991; Risch und Zhang, 1995). Daraus folgt, dass bei Verwendung unselektierter Stichproben der Großteil der Familien nur einen sehr geringen Beitrag zur Kopplungsanalyse liefert. Deshalb wäre es sinnvoll, Stichproben mit selektierten Familien zu verwenden, von denen zu erwarten ist, dass sie einen großen Beitrag zur Kopplungsanalyse leisten werden.

In den letzten 15 Jahren kam es infolge des gestiegenen Interesses an der genetischen Kartierung von Merkmalen mit quantitativer Ausprägung zu einer geradezu explosionsartigen Entwicklung kopplungsanalytischer Verfahren zur genetischen Kartierung quantitativer Phänotypen (siehe

Kapitel 2 oder Elston, 1998; Ferreira, 2004). Diese Verfahren verzichten auf die Annahme eines spezifischen genetischen Modells des Erbgangs (z.B. dominant oder rezessiv) und werden deshalb als modellfrei bezeichnet. Modellbasierte Verfahren sind demgegenüber aufgrund der Schwierigkeiten, bei quantitativen Merkmalen ein genetisches Vererbungsmodell herzuleiten, problematisch und deshalb nicht sehr verbreitet (siehe Kapitel 2, Abschnitt 8).

Die Literatur wird dabei immer wieder eine Klasse von Verfahren, nämlich die der VARIANZKOMPONENTENMODELLE (siehe Kapitel 2, Abschnitt 4), als Methode der Wahl erwähnt (siehe z.B. Blangero, 2004). Als Begründung wird die Möglichkeit der Schätzung der einzelnen Varianzkomponenten, sowie vor alledem die statistische Power im Vergleich zu anderen Methoden genannt. Eine wesentliche Voraussetzung für die Anwendung des Verfahrens ist jedoch die multivariate Normalverteilung des Phänotyps. Fehlt es an dieser Voraussetzung, so zeigen die VARIANZKOMPONENTENMODELLE z.T. massive Abweichungen vom nominalen Typ I-Fehler. Allison et al. (1999) haben z.B. für ein spezifisches genetisches Modell unter Verletzung der Normalverteilungsannahmen durch Monte-Carlo Simulationen einen empirischen Typ I-Fehler von 18% bei einem nominalen Fehler von 5% gezeigt. Der empirische Typ I-Fehler überschreitet also den nominalen Typ I-Fehler um mehr als 300%. Dies ist deswegen wichtig, weil als Studiendesign nicht nur RSP verwendet, sondern häufig auf selektierte Stichproben zurückgegriffen wird. Dieses Studiendesign verletzt in der Regel die Normalverteilungsannahmen (Dolan und Boomsma, 1998).

Nicht jedes der entwickelten Verfahren lässt sich also bei jedem Studiendesign oder Verteilungsform der Phänotypen verwenden. Diese Erkenntnis führt zu der Frage, nach welchen systematischen Kriterien die Güte eines kopplungsanalytischen Verfahrens zur genetischen Kartierung quantitativer Phänotypen beurteilt werden sollte.

Nach Feingold (2002) sollte hierfür von drei primären Kriterien ausgegangen werden. Das erste Kriterium ist die Power des Verfahrens unter idealen Bedingungen, wenn der nominale Typ I-Fehler korrekt gehalten wird. Für den Fall der quantitativen Phänotypen heißt dies, dass die Powervergleiche auf populationsbasierten (unselektierten) simulierten Stichproben basieren, bei denen der Phänotyp annähernd normal verteilt ist. Das zweite Kriterium ist die Robustheit des Typ I-Fehlers. Dabei geht es darum, ob sowohl unter Idealbedingungen als auch unter verschiedenen anderen Bedingungen - z.B. bei nicht-normal verteilten Phänotypen oder selektierte Stichproben das Typ I-Fehlerniveau korrekt gehalten wird. Als drittes Kriterium nennt Feingold die Robustheit der Power, d.h. die Frage, inwieweit z.B. nicht-normal verteilte Phänotypen oder selektierte Daten die Power eines Verfahrens beeinflussen. Zusätzlich zu den drei primären Kriterien soll nach Feingold der Einfluss abhängiger Geschwisterschaften auf den Typ I-Fehler zu beachten sein, d.h. ob Familien mit mehr als zwei Geschwistern rekrutiert wurden. Auch dies ist zu berücksichtigen, weil nur wenige Verfahren ursprünglich für die Anwendung auf abhängige Geschwisterschaften entwickelt oder fortentwickelt wurden.

Von den genannten Kriterien ausgehende Gütevergleiche zwischen verschieden kopplungsanalytischen Verfahren wurden bis jetzt nur in einem beschränkten Umfang durchgeführt.

Bei den meisten Gütevergleichen wurden dem zu beurteilenden Verfahren ein oder einige andere Verfahren auf der Basis von Monte-Carlo-Simulationen gegenübergestellt (Alcaïs und Abel, 1999; Allison et al., 2000; Sham und Purcell, 2001; Sham et al., 2002; Yu et al., 2004). Nur zwei auf Monte-Carlo Simulationen basierende Vergleiche wurden in der Klasse der regressionsbasierten Verfahren für das SPSP- und ESP-Design durchgeführt (Cuenco et al., 2003; Szatkiewicz et al., 2003). Für eine detaillierte Diskussion der Gütevergleiche wird an dieser Stelle auf die Vorstellung der einzelnen Verfahren in Kapitel 2 verwiesen.

Auch analytische Überlegungen wurden nur in einem beschränkten Umfang durchgeführt. Diesen algebraischen Vergleichen liegen in der Regel stark vereinfachte Annahmen zugrunde. So existieren theoretische Überlegungen zum Vergleich des HASEMAN-ELSTON Verfahrens (siehe Kapitel 2, Abschnitt 1) mit den VARIANZKOMPONENTENMODELLEN (Sham und Purcell, 2001) sowie zum Vergleich des HASEMAN-ELSTON Verfahrens und des WPC Verfahren (Commenges, 1994; Ziegler, 2001). Für das ESP Studiendesign wurde das MAXIMUM LIKELIHOOD BINOMIAL Verfahren (siehe Kapitel 2, Abschnitt 6) mit dem Verfahren von Risch und Zhang (1995) sowie dem EDAC Verfahren (Gu et al., 1996) verglichen. Da diese Gütevergleiche von unterschiedlichen Annahmen ausgehen, lassen sie sich nur schwer vergleichen oder kombinieren. Für einige Verfahren einschließlich des MERLIN-QTL Verfahrens (Abecasis et al., 2002, siehe Kapitel 2, Abschnitt 5) und des WILCOXON-RANGSUMMENTESTS (Kruglyak und Lander, 1995b, siehe Kapitel 2, Abschnitt 7) wurden bislang überhaupt keine Gütevergleiche durchgeführt. Darüber hinaus ist nur für einen Teil der Methoden bekannt, wie sich Abweichungen von der Normalverteilung bzw. die Analyse selektierter Stichproben auf Robustheit und Power auswirken.

In der Literatur wurde mehrfach darauf hingewiesen, dass weitere Studien erforderlich sind, um die Güteeigenschaften gerade im Vergleich zu anderen Verfahren sowie unter anderen Modellannahmen bzw. Studiendesigns zu ermitteln und zu vergleichen (Allison et al., 1999, S. 541; Allison et al., 2000, S. 252; Feingold, 2002, S. 220-221; Cuenco et al., 2003, S. 872; Szatkiewicz et al., 2003, S. 884).

Zusammenfassend lässt sich sagen, dass bislang ein umfassender Gütevergleich fehlt, der eine Vielzahl kopplungsanalytischer Verfahren zur genetischen Kartierung quantitativer Merkmale, unter verschiedenen für die Praxis wichtigen Bedingungen - wie Abweichungen von der Normalverteilung, verschiedene Studiendesigns (RSP, SPSP, ESP) sowie Einfluss abhängiger Geschwisterschaften - in einer Studie berücksichtigt.

In dem folgenden Abschnitt wird nun zunächst aufbauend auf diesem Abschnitt die Zielsetzung dieser Arbeit definiert. Im Kapitel 2 werden dann die wichtigsten kopplungsanalytischen Verfahren zur genetischen Kartierung von quantitativen Phänotypen im Detail erläutert und hergeleitet.

Ebenfalls werden die aus der Literatur bekannten Güteeigenschaften sowie wenn bekannt auch der Vergleich dieser mit den anderen Verfahren diskutiert.

## 1.2 Zielsetzung

Von den dargestellten Vorüberlegungen ausgehend, soll im Folgenden die Zielsetzung dieser Arbeit näher umrissen werden. In der Praxis werden zur Aufklärung komplexer genetischer Erkrankungen zunehmend kopplungsanalytische Methoden für quantitative Phänotypen unter Verwendung von Kernfamilien mit zwei oder mehr Geschwistern eingesetzt. In den letzten Jahren wurde eine Vielzahl neuer Verfahren für diese Fragestellungen entwickelt. Doch ist bisher weitgehend ungeklärt, wie sich die Güte dieser Verfahren im direkten Vergleich zueinander verhält.

Daher wird im Rahmen dieser Arbeit in einer Monte-Carlo Simulationsstudie die Güteeigenschaften einer Vielzahl verschiedener Verfahren unter verschiedenen Modellen und Studiendesigns verglichen.

Dabei werden acht Verfahren berücksichtigt, die in sechs für die nichtkommerzielle Nutzung freigegebenen Softwarepaketen verfügbar sind. Diese Verfahren werden unter drei genetischen Modellen (dominant, additiv, rezessiv) drei Studiendesigns (ohne Selektion [RSP], mit einfacher Selektion [SPSP] und doppelter Selektion [ESP]) und zwei Familienstrukturen (Kernfamilien mit einem Geschwisterpaar sowie Kernfamilien mit einer variierenden Anzahl von zwei bis fünf Geschwistern) untersucht. Zusätzlich wird der Effekt bei Abweichung von der Normalverteilung untersucht. Insgesamt werden also 36 verschiedene Simulations-Szenarien betrachtet. Hierbei werden die drei Kriterien nach Feingold (2002), im vorhergehenden Abschnitt näher beschrieben, zum Vergleich der Verfahren herangezogen:

1. Power unter Normalverteilungsannahmen, wenn das Fehlerniveau gehalten wird
2. Robustheit des Typ I-Fehlers gegenüber einer Verletzung der Normalverteilungsannahmen sowie unter verschiedenen Studiendesigns
3. Robustheit der Power

Zusätzlich wird, wie von Feingold empfohlen, der Einfluss abhängiger Geschwisterschaften betrachtet.

In einem ersten Schritt wird hierzu zunächst eine Simulationssoftware erstellt (`Sibsim`), anhand derer die Datensätze für die 36 Szenarien simuliert werden sollen. Durch eine externe Validierung werden dann unter verschiedenen Szenarien Datensätze erstellt und die einzelnen Simulationsparameter an zufällig ausgewählten Simulationen auf Abweichungen überprüft.

Für den Robustheitsvergleich werden dann für jedes Simulations-Szenario 100.000 Simulationen unter der Nullhypothese und zum Powervergleich 1.000 Simulationen unter der Alternativhypothese erstellt werden.

Der Robustheitsvergleich wird dabei durch Vergleich der Abweichungen zwischen den empirisch ermittelten Typ I-Fehleranteilen und dem nominalen Typ I-Fehler auf verschiedenen Testniveaus durchgeführt. Die hohe Anzahl der Simulationen unter der Nullhypothese ermöglicht dann auf Basis empirisch ermittelter Grenzwerte einen empirischen Powervergleich unter der Alternativhypothese, wie von Yu et al. (2004) zum Power-Vergleich von Verfahren zur genetischen Kartierung quantitativer Merkmale vorgeschlagen.

Die Anwendung der verschiedenen im Rahmen dieser Arbeit verwendeten Verfahren wird sodann an dem Datensatz „Consortium on Asthma Genetics: Perth study“ (COAG Perth Datensatz) illustriert (Palmer et al., 1998; Palmer et al., 2001).

# 2 Verfahren zur genetischen Kartierung quantitativer Merkmale

Im Folgenden sollen die wichtigsten kopplungsanalytischen Verfahren zur genetischen Kartierung quantitativer Phänotypen im Detail erläutert und hergeleitet werden. Daneben werden die aus der Literatur bekannten Güteeigenschaften diskutiert sowie – soweit verfügbar - auch der Vergleich zwischen diesen Verfahren und den anderen Verfahren analysiert.

Zunächst wird das genetische Modell zusammen mit dem HASEMAN-ELSTON Verfahren erläutert. Es folgen die Erweiterungen bis zum MERLIN-REGRESS Verfahren. Danach werden zunächst die VARIANZKOMPONENTENMODELLEN und die Allele-sharing Verfahren MERLIN-QTL und das MAXIMUM-LIKELIHOOD-BINOMIAL Verfahren erläutert. Abschließend werden der nichtparametrische WILLCOXON-RANGSUMMENTEST sowie die MODELLBASIERTE Kopplungsanalyse erläutert.

## 2.1 Haseman-Elston Verfahren

Im Jahr 1972 stellten Haseman und Elston ein modellfreies kopplungsanalytisches Verfahren zur genetischen Kartierung quantitativer Phänotypen auf der Basis eines Regressionsmodells vor (Haseman und Elston, 1972). Es stellt den Ausgangspunkt für die genetische Kartierung quantitative Phänotypen dar und ist eine der am häufigsten zitierten Arbeiten im Zusammenhang mit der Kopplungsanalyse quantitativer Phänotypen. Das HASEMAN-ELSTON Verfahren wird wegen seiner Einfachheit auch heute noch vielfach angewendet.

Ihm liegt folgender Gedanke zugrunde: Ähneln zwei Geschwister einander phänotypisch und wird die Ausprägung des Phänotyps dabei maßgeblich durch einen genetischen Locus (der im folgenden QTL genannt wird) beeinflusst, dann sollten sich die beiden Personen an diesem Locus auch genetisch ähneln. Das Verfahren verlangt daher, dass zunächst Maße für die genetische und die phänotypische Ähnlichkeit definiert werden.

Für die genetische Ähnlichkeit des Geschwisterpaars $m$ in einem Stammbaum können die Allele identical-by-descent (IBD) als Maßstab herangezogen werden. Dies meint die Anzahl der Allele, die zwei Personen in einem Stammbaum aus gleicher Herkunft gemeinsam vererbt wurden. Für Geschwister kann der IBD-Wert 0, 1 oder 2 sein. Beim HASEMAN-ELSTON Verfahren wird jedoch der Anteil der Allele IBD $\tau$ betrachtet. Für die IBD-Werte 0, 1 oder 2 ergeben sich dann für den Anteil der Allele IBD die Werte 0, ½, 1.

Als Maß für die phänotypische Ähnlichkeit verwendet das HASEMAN-ELSTON Verfahren die quadrierte phänotypische Differenz $y$, also den euklidischen Abstand.

Zur Herleitung des Regressionsmodells betrachten Haseman und Elston ein einfaches additives Modell. Wenn mit $x_{1m}$ und $x_{2m}$ die beobachteten Phänotypen des $m_{\text{ten}}$ Geschwisterpaares bezeichnet werden, dann ist das additive Modell gegeben durch:

$$x_{1m} = \mu + g_{1m} + \varepsilon_{1m}$$
$$x_{2m} = \mu + g_{2m} + \varepsilon_{2m}$$

Wobei mit $\mu$ der allgemeine Mittelwert, mit $g_{im}$ der Hauptgeneffekt und mit $\varepsilon_{im}$ die Residualgröße der Person $i$ vom $m_{\text{ten}}$ Geschwisterpaar bezeichnet wird. Polygene Effekte und Umwelteffekte gehen in die Residualgröße $\varepsilon_{im}$ ein. Der Hauptgeneffekt wird dabei von einem biallelischen Locus mit den Allelen $A_1$ und $A_2$ bestimmt.

Unter der Annahme, dass kein Dominanzeffekt vorliegt, ergibt sich das von Haseman und Elston (1972) vorgeschlagene Regressionsmodell dann wie folgt:

$$y_m = \alpha + \beta\tau_m$$

wobei $y_m$ die quadrierte phänotypische Differenz ist und die Regressionskoeffizienten $\alpha$ und $\beta$ gegeben sind durch:

$$\alpha = \sigma_\varepsilon^2 + 2\sigma_g^2$$
$$\beta = -2\sigma_g^2$$

Wenn also $\hat{\beta}$ ein Schätzer für $\beta$ ist, dann ist $-1/2\,\hat{\beta}$ ein Schätzer für $2\sigma_g^2$.

Wenn die elterlichen Genotypen bestimmt sind und mit in die Analyse eingehen, dann führt – wie Amos et al. (1989) gezeigt haben – die Vernachlässigung eines möglichen Dominanzterms (siehe Kapitel 3, Abschnitt 1.1) nicht zu einem verzerrten Schätzer. Werden die elterlichen Genotypen nicht bestimmt und ist eine Dominanzkomponente vorhanden, so ist $\hat{\beta}$ ein verzerrter Schätzer für $\beta$. Im Allgemeinen ist dieser Bias aber vernachlässigbar (Amos et al., 1990). In der Realität wird deshalb ein möglicher Dominanzterm meist vernachlässigt.

Kopplung zwischen einem Markerlocus und einem quantitativen Phänotyp liegt vor, wenn der geschätzte Regressionskoeffizient $\hat{\beta}$ signifikant kleiner als 0 ist. Für den Fall, dass der Markerlocus und der quantitative Phänotyp nicht gekoppelt sind, ist $\beta = 0$. Der statistische Test auf Kopplung ist also einseitiger $t$-Test auf den Parameter $\beta$.

Das HASEMAN-ELSTON Verfahren ist zunächst für Kernfamilien mit einem Geschwisterpaar hergeleitet worden. Bei einer größeren Anzahl von Geschwistern sind die Geschwisterpaare jedoch nicht mehr unabhängig; dies kann dazu führen, dass der $p$-Wert überschätzt wird. Das HASEMAN-ELSTON Verfahren tendiert also bei einer größeren Anzahl von Geschwistern dazu, liberal zu werden (siehe z.B. Williams und Blangero, 1999). Zwei mögliche Lösungsansätze sollen hier kurz vorgestellt werden. Zum einen kann man nur die strikt unabhängigen Geschwisterpaare berücksichtigen. Damit wird aber ein Teil der vorhandenen Informationen vernachlässigt, was zwangsläufig zu einem Powerverlust führt. Eine weitere Möglichkeit besteht darin, eine generalisierte Kleinst-Quadrate-Regression zu verwenden, die eine Korrelation zwischen den quadrierten phänotypischen Differenzen erlaubt. Damit ist es möglich, alle möglichen Geschwisterpaare und

damit die kompletten zur Verfügung stehenden Informationen zu verwenden. Zuerst wurde diese Idee von Single und Finch (1995) beschrieben. Die Autoren haben gezeigt, dass bei mehr als zwei Geschwistern die Verwendung einer generalisierten Kleinst-Quadrate-Regression im Vergleich zur Analyse der unabhängigen Geschwisterpaare zu einem deutlichen Anstieg der Power führt.

Elston et al. haben bei der Vorstellung des REVIDIERTEN HASEMAN-ELSTON Verfahrens (siehe Kapitel 2, Abschnitt 2) diese Idee für die Anwendung auf weitere Kovariaten und multiple QTL's erweitert und diesen Ansatz in das Softwarepaket `S.A.G.E.` implementiert (Elston et al., 2000).

Das HASEMAN-ELSTON Verfahren ist in einer Vielzahl von Softwarepaketen implementiert, neben dem soeben erwähnten `S.A.G.E.` beispielsweise auch in dem gleichfalls häufig verwendeten Programm `Genehunter`.

## 2.2 Revidiertes Haseman-Elston Verfahren

Das klassische HASEMAN-ELSTON Verfahren wurde im Laufe der Zeit vielfach modifiziert und erweitert. Das ursprüngliche Verfahren wurde vor allem dafür kritisiert, dass infolge der Verwendung der quadrierten phänotypischen Differenz nicht die gesamten in den Daten vorhandenen Informationen ausgenutzt werden (Fulker und Cherny, 1996; Wright, 1997; Drigalenko, 1998). Wright zeigte 1997, dass ein nicht zu unterschätzender Gewinn an Power erreicht werden kann, wenn auch die Information der phänotypischen Summe verwendet wird.

Dies führte unter anderem zu dem REVIDIERTEN HASEMAN-ELSTON Verfahren wie es von Elston im Jahre 2000 vorgeschlagen wurde (Elston et al., 2000).

Ausgehend von der quadrierten Differenz der zentrierten Phänotypen des $m_{\text{ten}}$ Geschwisterpaares

$$y_{m,D} = \left( \left( x_{1m} - \mu \right) - \left( x_{2m} - \mu \right) \right)^2$$

und der quadrierten Summe der zentrierten Phänotypen, nämlich

$$y_{m,S} = \left( \left( x_{1m} - \mu \right) + \left( x_{2m} - \mu \right) \right)^2$$

wie von Wright (1997) vorgeschlagen wurde, können diese beiden Informationen durch die Betrachtung der Differenz der beiden Größen $y_{m,S} - y_{m,D}$ kombiniert werden.

Die Differenz $y_{m,S} - y_{m,D}$ ist dabei identisch mit dem 4-fachem des zentrierten Kreuzprodukts

$$y_{m,S} - y_{m,D} = 4\left( x_{1m} - \mu \right)\left( x_{2m} - \mu \right),$$

welche im REVIDIERTEN HASEMAN-ELSTON Verfahren als abhängige Variable für die Regression verwendet wird.

Elston (2000) hat durch Simulationen gezeigt, dass bei Familien mit zwei Kindern das Typ I-Fehlerniveau besser kontrolliert werden kann.

Auch unter Verletzung der Normalverteilungsannahmen der Phänotypen, bei starker Residualkorrelation der Geschwister und unter der Bedingung selektierter Familien, bei denen ein Geschwister

aus dem untersten und oder obersten Dezil der Verteilung stammt, wird das Typ I-Fehlerniveau korrekt gehalten (Allison et al., 2000).

Das REVIDIERTE HASEMAN-ELSTON Verfahren hat im Vergleich zum ursprünglichen HASEMAN-ELSTON Verfahren eine größere Power, wenn die Korrelation der Geschwister klein ist, aber eine geringere Power, wenn die Korrelation der Geschwister groß ist (Palmer et al., 2000; Forrest, 2001).

Das REVIDIERTE HASEMAN-ELSTON Verfahren ist in dem Softwarepaket `S.A.G.E.` implementiert.

## 2.3 Merlin-Regress Verfahren

Im Jahre 2002 stellten Sham et al. (2002) ein neues regressionbasiertes Verfahren zur genetischen Kartierung quantitativer Phänotypen vor (im Folgenden mit MERLIN-REGRESS bezeichnet). Grundgedanke dieses Verfahrens ist, das HASEMAN-ELSTON Verfahren umzukehren. Darüber hinaus verwenden die Autoren als Maß für die phänotypische Ähnlichkeit nicht nur die quadrierte Differenz, sondern auch die quadrierte Summe in einer multivariaten Regression. Ein weiterer wesentlicher Fortschritt dieser Methode im Vergleich zum HASEMAN-ELSTON bzw. dem REVIDIERTEN HASEMAN-ELSTON Verfahren ist, das nicht nur Geschwisterpaare, sondern Paare aller Verwandtschaftsgrade in die Berechnung der Teststatistik mit einbezogen werden können.

Die Autoren haben durch Simulationen gezeigt, dass ihr Verfahren das Typ I-Fehlerniveau sowohl unter Normalverteilungsannahmen als auch unter Verletzung der Normalverteilungsannahmen und bei Familien mit abhängigen Geschwisterschaften hält. Darüber hinaus haben Sham et al. zusätzlich gezeigt, dass das Typ I-Fehler-Niveau auch bei konkordanten oder diskordanten ESP-Studiendesigns gehalten wird. MERLIN-REGRESS ist somit scheinbar sehr robust und auch auf selektierte Datensätze und nicht-normal verteilte Phänotypen anwendbar. Die Power soll nach Aussage der Autoren vergleichbar mit der hohen Power der Varianzkomponentenmodelle sein.

Im Gegensatz zu den anderen hier vorgestellten Methoden setzt die Anwendung von MERLIN-REGRESS jedoch Schätzungen des populationsbasierten Mittelwerts, der Varianz sowie der Heritabilität voraus. Die Schätzung dieser Parameter kann jedoch im Falle von selektierter Datensätze bzw. bei nicht normal verteilten Phänotypen schwierig sein und zu verzerrten Schätzungen führen. Für den Fall von unselektierter Datensätze und normal verteilter Phänotypen haben Sham et al. durch Simulationen gezeigt, dass falsche Parameter lediglich zu einem Powerverlust führen und keinen Einfluss auf den Typ I-Fehler haben. Ob sich diese Aussagen auch auf selektierte Familien und/oder auf nicht normal verteilten Phänotypen übertragen lassen, ist jedoch noch unklar.

Zur Herleitung des Verfahrens werden zunächst zwei Vektoren **S** und **D** definiert, welche die quadrierten Phänotypsummen $y_{jk,S}$ und die quadrierten Phänotypdifferenzen $y_{jk,D}$ für alle Paare der

Personen $j$ und $k$ mit $j \neq k$ eines Stammbaumes enthalten. Darüber hinaus sei $\hat{\mathbf{\Pi}}$ ein Vektor, der die Schätzer für den zentrierten Anteil der Allele IBD $\hat{\tau}_{jk}$ aller Paare der Personen $j$ und $k$ mit $j \neq k$ enthält.

Für Familien mit mehr als vier Mitgliedern ergeben sich jedoch Kolinearitäten zwischen **S** und **D**. Um diese Kolinearitäten zu entfernen, wird deshalb der Vektor **D** willkürlich auf die Anzahl der Familienmitglieder gekürzt, wobei jedes Individuum mindestens einmal vorkommt. Da die entfallenen Elemente von **D** lineare Kombinationen der beibehaltenen Elemente von **S** und **D** sind, resultiert aus der Kürzung gemäß Sham et al. (2002) kein Informationsverlust. Der gekürzte Vektor von **D** wird mit $\mathbf{D}^*$ bezeichnet. Die beiden Vektoren **S** und $\mathbf{D}^*$ werden dann zu dem Vektor $\mathbf{Y_{MR}} = [\mathbf{S}, \mathbf{D}^*]'$ zusammengefasst. Damit das MERLIN-REGRESS Verfahren auch auf selektierte Stichproben anwendbar ist, werden $\mathbf{Y_{MR}}$ und $\hat{\mathbf{\Pi}}$ zentriert:

$$\mathbf{Y_{C,MR}} = \mathbf{Y_{MR}} - E\left(\mathbf{Y_{MR}}\right)$$

$$\hat{\mathbf{\Pi}}_{\mathbf{C}} = \hat{\mathbf{\Pi}} - E\left(\hat{\mathbf{\Pi}}\right)$$

Die multivariate Regression von $\hat{\mathbf{\Pi}}_{\mathbf{C}}$ auf $\mathbf{Y_{C,MR}}$ ist dann gegeben durch

$$\hat{\mathbf{\Pi}}_{\mathbf{C}} = \mathbf{\Sigma}'_{\mathbf{Y_{MR}}\hat{\mathbf{\Pi}}} \mathbf{\Sigma}^{-1}_{\mathbf{Y_{MR}}} \mathbf{Y_{C,MR}} + \boldsymbol{\varepsilon},$$

wobei $\mathbf{\Sigma}_{\mathbf{Y_{MR}}\hat{\mathbf{\Pi}}}$ die Kovarianzmatrix zwischen $\mathbf{Y_{MR}}$ und $\hat{\mathbf{\Pi}}$, und $\Sigma_{\mathbf{Y_{MR}}}$ die Kovarianzmatrix von $\mathbf{Y_{MR}}$ ist. $\boldsymbol{\varepsilon}$ bezeichnet das Residuum.

Die Matrix $\mathbf{\Sigma}_{\mathbf{Y_{MR}}\hat{\mathbf{\Pi}}}$ kann faktorisiert werden in $\mathbf{Q}\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}\mathbf{H}$, wobei **Q** eine Diagonalmatrix mit den Werten $\sigma_g^2$ ist. Die Matrix **H** ist dann eine horizontale Blockmatrix, wobei der erste Block eine quadratische Diagonalmatrix mit den konstanten Werten 2 und die zweite Blockmatrix eine Diagonalmatrix mit den konstanten Werten -2 ist. $\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}$ ist die Kovarianzmatrix der geschätzten zentrierten Anteile IBD. Die Schätzung von $\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}$ sowie $\mathbf{\Sigma}_{\mathbf{Y_{MR}}}$ ist gegeben in Sham et al. (2002) und wird hier nicht weiter im Detail vorgestellt.

Wenn $\mathbf{H}\mathbf{\Sigma}^{-1}_{\mathbf{Y}}\mathbf{Y_{MR}}$ im Folgenden mit **B** bezeichnet wird, dann ist nach Sham et al. der optimal gewichtete Schätzer für $\sigma_g^2$ pro Familie durch

$$\frac{\mathbf{B}'\hat{\mathbf{\Pi}}_{\mathbf{C}}}{\mathbf{B}'\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}\mathbf{B}}$$

bzw. für alle Familien einer Stichprobe durch

$\hat{\sigma}_g^2 = \dfrac{\sum[\mathbf{B}'\hat{\mathbf{\Pi}}_{\mathbf{C}}]}{\sum[\mathbf{B}'\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}\mathbf{B}]}$ gegeben.

Die Teststatistik lässt sich dann wie folgt formulieren:

$$T = \hat{\sigma}_g^2 \sum [\mathbf{B}' \hat{\mathbf{\Pi}}_C]$$

Unter der Nullhypothese ist $T$ $\chi^2$ verteilt mit einem Freiheitsgrad. Da nur positive Werte von $\sigma_g^2$ Sinn machen, empfehlen die Autoren, $T$ auf null zu setzen, wenn $\hat{\sigma}_g^2$ negativ ist. Die Teststatistik folgt dann unter der Nullhypothese einer 50:50 Mischung aus 0 und einer $\chi^2$ Verteilung mit einem Freiheitsgrad.

Das Verfahren von Sham et al. ist in dem Softwarepaket `Merlin` implementiert.

## 2.4 Varianzkomponentenmodelle

Mitte der $90_{er}$ Jahre boten die VARIANZKOMPONENTENMODELLE erstmals eine wichtige Alternative zum HASEMAN-ELSTON Verfahren (Amos, 1994; Almasy und Blangero, 1998).

Die Varianzkomponentenmodelle basieren auf einer Erweiterung des additiven Modells, wie es zuvor zur Herleitung des HASEMAN-ELSTON-Verfahrens verwendet wurde:

$$x_m = \mu + g_m + G_m + \beta_1' u_m + \varepsilon_m$$

Das additive Modell ist hier um zwei zusätzliche Terme erweitert. $G_m$ ist ein zufälliger polygener Effekt des $m_{\text{ten}}$ Geschwisterpaares. $G_m$ wird also nicht wie beim HASEMAN-ELSTON Verfahrens in der Residualgröße sondern als eigener Term betrachtet. Zusätzlich zu den genetischen Variablen können $p$ Kovariablen, die in einem $p \times 1$ Vektor $u_m$ zusammengefasst werden in das Modell aufgenommen werden. Der Effekt der Kovariablen wird durch den $p \times 1$ Parametervektor $\beta_l$ beschrieben.

Die Varianz der Phänotypen ist aufgrund des additiven Modells gegeben durch die Summe $\sigma_a^2 + \sigma_d^2 + \sigma_G^2 + \sigma_\varepsilon^2$,

wobei $\sigma_a^2$ die additive Varianz und $\sigma_d^2$ die Dominanzvarianz des Hauptgeneffekts ist.

Die Schätzung der vier Varianzkomponenten $\sigma_a^2$, $\sigma_G^2$, $\sigma_\varepsilon^2$ und $\sigma_d^2$ kann unter Verwendung der Maximum-Likelihood Methode durchgeführt werden. Der statistische Test auf Kopplung wird als LQT unter einem unbeschränkten Modell (in dem $\sigma_a^2$, $\sigma_G^2$, $\sigma_\varepsilon^2$ und $\sigma_d^2$ und ein mögliches $\theta$ geschätzt werden) und einem beschränkten Modell unter der Nebenbedingung $\sigma_a^2 = 0$ durchgeführt:

$$LQT = \frac{L(\sigma_a^2 = 0, \hat{\sigma}_G^2, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_d^2, \hat{\theta})}{L(\hat{\sigma}_a^2, \hat{\sigma}_G^2, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_d^2, \hat{\theta})}$$

Gemäß klassischer LQT-Theorie ist -2ln LQT asymptotisch $\chi^2$-verteilt mit einem Freiheitsgrad.

Während das HASEMAN-ELSTON Verfahren in seiner ursprünglichen Form nur auf Geschwister anwendbar ist, können die Varianzkomponentenmodelle für beliebige Stammbäume verwendet werden. Ein weiterer Vorteil liegt darin, dass die Varianzkomponentenmodelle es nicht nur erlauben, einen Test auf Kopplung durchzuführen, sondern zusätzlich auch die einzelnen Varianzkomponenten zu schätzen.

Zahlreiche Simulationsstudien haben gezeigt, dass die Varianzkomponentenmodelle eine wesentlich höhere Power als das HASEMAN-ELSTON Verfahren haben (siehe z.B. Amos et al., 1996). Sie zeigen jedoch bei Verletzung der zugrunde liegenden multivariaten Normalverteilungsannahmen teilweise einen inakzeptabel liberalen Typ I-Fehler (siehe z.B. Allison et al., 1999). Blangero et al. vermuten, dass sich ein liberaler Typ I-Fehler meist auf eine positive Wölbung der Phänotypverteilung (4. Moment der Verteilung) zurückführen lässt (Blangero et al., 2001). Wenn die Wölbung der Phänotypverteilung größer als 1,5 ist, dann empfehlen Blangero et al. alternative robustere Methoden zu verwenden. Als Alternative schlägt Blangero z.B. die Schätzung der Varianzkomponenten durch robuste Methoden wie die Pseudo-Maximum-Likelihood Schätzung vor (Amos, 1994).

Die zugrunde liegenden multivariaten Normalverteilungsannahmen werden - wie bereits in der Einleitung erwähnt – auch durch die Analyse selektierter Datensätze verletzt (Dolan und Boomsma, 1998). In der Praxis werden die VARIANZKOMPONENTENMODELLE mit selektierten Datensätzen jedoch unzulässigerweise häufig ohne Berücksichtigung dieses Umstandes verwendet.

Die Varianzkomponentenmodelle sind unter anderem in den Softwarepaketen `Genehunter`, `Merlin` und `Solar` implementiert.

## 2.5 Merlin-QTL Verfahren

Grundlage des MERLIN-QTL Verfahrens sind die beiden Arbeiten von Whittemore und Halpern (1994) und Kong und Cox (1997) zu Allele-sharing Statistiken, dessen Ziel allerdings die Analyse binärer Phänotypen ist. Wie von Ferreira (2004) beschrieben, wurde dieses Verfahren jedoch von Abecasis (2002) für die Verwendung von quantitativen Phänotypen angepasst.

MERLIN-QTL definiert zur genetischen Kartierung quantitativer Phänotypen eine Funktion $S(w)$, welche jeden möglichen Vererbungsvektor $w$ für einen gegebenen Stammbaum hinsichtlich seiner Evidenz für Kopplung bewertet. Je höher die Bewertung von $S$ für einen gegebenen Vererbungsvektor $w$ ist, desto größer ist die Evidenz für Kopplung für diesen Vererbungsvektor.

Die von MERLIN-QTL verwendete Scoring-Funktion $S(w)$ ist

$$S(w) = \sum_a S_a^{\,2} ,$$

mit $S_a = \sum_i (y_i - \mu)$.

Die Bewertung für jeden Vererbungsvektor *w* in einer Familie wird also durch Summation der quadrierten Bewertungen von allen Gründer-Allelen *a* (englisch: founder alleles) berechnet, die in dem Vererbungsvektor vorhanden sind. Die Bewertung für jedes Gründer-Allel ($S_a$) in dem Vererbungsvektor *w* wird durch die Summation der Abweichungen der Phänotypen vom populationsbasierten Mittelwert für alle Personen *i* berechnet, die dieses Gründer-Allel tragen.

Basierend auf der Arbeit von Whittemore und Halpern (1994) werden dann die Bewertungen aller Stammbäume in *Z*-Scores umgerechnet. Zusätzlich wird, wie von Kong und Cox (1997) vorgeschlagen, ein LQT formuliert.

Das MERLIN-QTL Verfahren ist in dem Softwarepaket `Merlin` implementiert.

## 2.6 Maximum Likelihood Binomial Verfahren

Das MAXIMUM LIKELIHOOD BINOMIAL Verfahren für quantitative Phänotypen (Alcaïs und Abel, 1999) ist eine Erweiterung des MAXIMUM LIKELIHOOD BINOMIAL Verfahren für binäre Phänotypen (Abel et al., 1998; Abel und Müller-Myhsok, 1998).

Das MAXIMUM LIKELIHOOD BINOMIAL Verfahren basiert dabei auf der Idee der binomial verteilten Weitergabe der elterlichen Allele auf die Nachkömmlinge. Liegt keine Kopplung vor, erfolgt keine präferentielle Transmission eines Allels. Sind Marker und Krankheit hingegen gekoppelt, wird eine von 0,5 abweichende Transmission erwartet.

Wir betrachten eine Kernfamilie mit *n* Geschwistern. $\mathbf{X} = (x_1, x_2, ..., x_n)'$ sei der $n \times 1$ Vektor der Phänotypen und $\mathbf{M} = (m_{11}, m_{12}, ..., m_{n1}, m_{n2})'$ sei der $2n \times 1$ Vektor der Allele der Geschwister am Markerlocus.

Zur Konstruktion eines LQT benötigen wir zunächst die Likelihood der beobachteten Marker-Allele gegeben den Phänotypen der Kinder: $P(\boldsymbol{M}|\boldsymbol{X})$

Dazu wird zunächst eine latente binäre Variable eingeführt, welche die Kopplungsinformationen zwischen dem QTL und dem Marker enthält. Deshalb sei $\mathbf{B} = (b_1, b_2, ..., b_n)'$ der Vektor dieser binären Variablen für die Geschwister einer Familie. Da $b_i$ nicht beobachtbar ist, wird die Einführung von **B** in $P(\boldsymbol{M}|\boldsymbol{X})$ durch Summation der $2^n$ möglichen **B** Vektoren erreicht, denn per Definition ist **M** und **X**, gegeben **B**, bedingt unabhängig.

$$P(\mathbf{M} \mid \mathbf{X}) = \sum_n P(\mathbf{B} \mid \mathbf{X}) * P(\mathbf{M} \mid \mathbf{B}, \mathbf{X}) = \sum_n P(\mathbf{B} \mid \mathbf{X}) \cdot P(\mathbf{M} \mid \mathbf{B})$$

Die Formulierung der Likelihood $P(\mathbf{M}|\mathbf{X})$ setzt also $P(\mathbf{B}|\mathbf{X})$ und $P(\mathbf{M}|\mathbf{B})$ voraus.

Formulierung von $P(\mathbf{B}|\mathbf{X})$

Da die latente binäre Variable des *i*-ten Geschwisters $b_i$ nur von $x_i$ abhängt, ist die gemeinsame bedingte Verteilung von **B** gegeben **X** das Produkt der univariaten Verteilungen:

$$P(\mathbf{B} \mid \mathbf{X}) = \prod_{i=1}^{n} P(b_i \mid x_i)$$

Bei der Definition von $P(b_i|x_i)$ ist von folgender Überlegung auszugehen: Je höher der Wert von $x_i$ ist, desto größer sollte $P(b_i = 1|x_i)$ sein. Theoretisch lässt sich jede beliebige Verteilungsfunktion als Verbindungsfunktion wählen. Das MAXIMUM LIKELIHOOD BINOMIAL Verfahren lässt sich deshalb sowohl unter der Annahme einer Verteilung (klassischer Weise unter Annahme einer Normalverteilung) als auch ohne eine Annahme zur Art der Verteilung verwenden. In zweiten Fall wird die Verteilung dann durch eine empirische Verteilungsfunktion als Treppenfunktion über die kumulative Häufigkeiten definiert (z.B. durch Verwendung der populationsbasierten empirischen Dezile).

Formulierung von $P(\mathbf{M}|\mathbf{B})$

Die Definition von $P(\mathbf{M}|\mathbf{B})$ basiert auf dem Gedanken, dass die Vererbung der elterlichen Allele auf die Kinder einer Binomialverteilung folgt.

Unter der Nullhypothese (keine Kopplung) erbt jedes Geschwister mit einer Wahrscheinlichkeit von 0,5 das Marker Allel *A* (oder *B*) von den heterozygoten Eltern mit dem Genotyp *AB*.

*S* sei die Anzahl der Geschwister mit $b_i = 0$ für den gegeben Vektor **B**. Weiterhin sei $\alpha$ die Wahrscheinlichkeit, dass Geschwister mit $b_i = 1$ das Allel *A* von einem Elternteil mit dem Genotyp *AB* erhalten haben. Dementsprechend ist $1\text{-}\alpha$ die Wahrscheinlichkeit für Geschwister mit $b_i = 0$. Wenn außerdem die Elternteile mit *j* bezeichnet werden, dann ist die Likelihood der Familie

$$P(\mathbf{M} \mid \mathbf{B}) = \prod_{j=1}^{2} g_j(\alpha),$$

wobei $g_j(\alpha)$ für die Likelihood eines Geschwisters für den Elternteil *j* steht. Der Beitrag zur Likelihood $g_j(\alpha)$ eines Geschwisters mit $b_i = k$ ist dann gegeben durch $\alpha^k(1 - \alpha)^{1-k}$ oder $\alpha^{1-k}(1 - \alpha)^k$, wenn dem Geschwister das Allel *A* oder *B* vererbt wurde. Für den Fall, dass die Phase der Vererbung nicht bekannt ist, wissen wir von einer Familie zur nächsten nicht, welches Marker-Allel bei der Vererbung die Wahrscheinlichkeit $\alpha$ und welches die Wahrscheinlichkeit $1\text{-}\alpha$ hat. Jede dieser Möglichkeiten hat die Prätestwahrscheinlichkeit von 0,5. Deshalb kann $g_j(\alpha)$ dann wie folgt formuliert werden:

$$g_j(\alpha) = 0{,}5\left[\alpha^{n_a(1)}(1-\alpha)^{S-n_A(1)}\right]\left[\alpha^{n-S-n_A(0)}(1-\alpha)^{n_A(0)}\right] + 0{,}5\left[(1-\alpha)^{n_a(1)}\alpha^{S-n_A(1)}\right]\left[(1-\alpha)^{n-S-n_A(0)}\alpha^{n_A(0)}\right]$$

Hierbei ist $n_A(k)$ die Anzahl der Geschwister mit $b_i = k$, denen das Allel *A* von den heterozygoten Eltern mit den Genotypen *AB* vererbt wurde.

Jetzt lässt sich die die Likelihood-Funktion für die Familie *f* mit zwei Eltern und *n* Kindern wie folgt formulieren:

$$L_f(\alpha) = \sum_n \prod_{i=1}^{n} P(b_i \mid x_i) \prod_{j=1}^{2} g_j(\alpha)$$

Die Likelihood des gesamten Datensatzes $L(\alpha)$ mit $F$ Familien ist dann das Produkt über die $F$ Familien von $L_f(\alpha)$.

Sei $\hat{a}$ der Maximum-Likelihood Schätzer für $\alpha$, dann ist der Test auf Kopplung als LQT gegeben durch

$$\lambda_{mlb} = \frac{L(\alpha = 0{,}5)}{L(\alpha = \hat{a})}.$$

Entsprechend klassischer LQT Theorie ist -2ln $\lambda_{mlb}$ asymptotisch $\chi^2$-verteilt mit einem Freiheitsgrad. Der Test sollte grundsätzlich einseitig ausgeführt werden.

Für Familien mit unterschiedlicher Anzahl von Geschwisterschaften haben Alcaïs und Abel (1999) durch Simulationen gezeigt, das sowohl das Typ I-Fehler-Niveau gehalten wird als auch das MAXIMUM LIKELIHOOD BINOMIAL Verfahren generell eine höhere Power als das HASEMAN-ELSTON Verfahren zeigt.

Durch die Verwendung einer empirischen Verteilungsfunktion lässt sich das MAXIMUM LIKELIHOOD BINOMIAL Verfahren auch für selektierte Familien als auch nicht normalverteilter Phänotypen adäquat anwenden (Alcaïs und Abel, 1999). Für extrem diskordanten Geschwisterschaften haben Alcaïs und Abel analytisch gezeigt, dass das MAXIMUM LIKELIHOOD BINOMIAL Verfahren eine höhere Power als der vorgeschlagene Test von Risch und Zhang (1995) hat. In Daten, die sowohl extrem konkordante als auch diskordante Geschwister-Pärchen enthalten, haben Alcaïs und Abel durch Simulationen gezeigt, dass das MAXIMUM LIKELIHOOD BINOMIAL Verfahren zumindest die gleiche statistische Power wie die EDAC-Methode von Gu et al. (1996) hat.

Das MAXIMUM LIKELIHOOD BINOMIAL Verfahren wurde in der Vergangenheit im Vergleich zu den populären Methoden wie die VARIANZKOMPONENTENMODELLE oder das HASEMAN-ELSTON Verfahren relativ selten verwendet. Knobloch et al. (2000) verwendete das MAXIMUM LIKELIHOOD BINOMIAL Verfahren, um auf dem langen Arm von Chromosom 13 einen QTL zu kartieren, der die LDL-Serumkonzentration beeinflusst. Dina et al. haben 2005 auf dem kurzen Arm von Chromosom 8 einen QTL zur habituellen Ängstlichkeit kartiert (Dina et al., 2005).

Zur Verwendung des MLB QT ist derzeit nur das das Softwarepaket `Mlbgh` verfügbar. Dabei handelt es sich um eine Modifikation des Softwarepaketes `Genehunter`.

## 2.7 Wilcoxon-Rangsummentest

Kruglyak und Lander stellten 1995 ein modellfreies nichtparametrisches Kopplungsverfahren für quantitative Phänotypen auf der Basis eines WILCOXON-RANGSUMMENTEST vor. Der WILCOXON-

RANGSUMMENTEST ist auf beliebige Verteilungen der Phänotypen anwendbar und wird deshalb von den Autoren besonders für die Kopplungsanalysen bei nicht normal verteilten Phänotypen vorgeschlagen (Kruglyak und Lander, 1995b).

Ausgangspunkt ist wie beim Verfahren von HASEMAN UND ELSTON die quadrierte phänotypische Differenz von $n$ Geschwisterpaaren. Diese werden im ersten Schritt entsprechend ihrer Ränge geordnet, wobei $r_m$ der Rang des $m$-ten Geschwisterpaars ist. Weiterhin werden die IBD-Werte über die Funktion $f$ zentriert: $f$ nimmt die Werte -1, 0 bzw. 1 an, wenn die IBD-Werte 0, 1 oder 2 sind. Die von Kruglyak und Lander vorgeschlagene Rangsummenstatistik ist dann gegeben durch

$$T = \sum_{m=1}^{n} r_m \cdot f(m)$$

Aufgrund des zentralen Grenzwertsatzes ist $T$ asymptotisch normalverteilt, unter $H_0$ mit Erwartungswert $E(T) = 0$ und Varianz $\mathrm{Var}(T) = \frac{n(n+1)(2n+1)}{12}$. Bei einer hinreichenden Anzahl von Geschwisterpaaren lässt sich also ein einseitiger asymptotischer $z$-Test anwenden.

Der WILCOXON-RANGSUMMENTEST ist in den Programmen `Mapmaker/Sibs` (Kruglyak und Lander, 1995a) und `Genehunter` (Kruglyak et al., 1996) implementiert.

## 2.8 Modellbasiertes Verfahren

Neben den modellfreien Verfahren gibt es auch die Möglichkeit eine voll parametrisierte LQT Analyse als MODELLBASIERTES Verfahren zu verwenden (Lathrop et al., 1984). In der Praxis sind MODELLBASIERTE kopplungsanalytische Verfahren zur genetischen Kartierung quantitativer Phänotypen nicht weit verbreitet. Dieses liegt primär daran, dass die Verwendung eine detaillierte Modellspezifikation erfordert. So setzt die modellbasierte Berechnung die Annahme eines genetischen Vererbungsmodells und eine genaue Spezifikation der Vererbungsparameter voraus. Für den angenommenen QTL müssen dann die Allelfrequenzen sowie die dazugehörigen Erwartungswerte des Hauptgeneffektes zusammen mit der Varianz des Phänotyps definiert werden. Da für diese Parametrisierung keine sinnvollen Schätzmethoden zur Verfügung stehen und die Ergebnisse bei willkürlich gesetzten Parametern kaum interpretierbar sind, wird daher in Anwendungen üblicherweise modellfreien kopplungsanalytischen Verfahren der Vorzug gegeben.

Das modellbasiertes Verfahren ist im `Linkage` Softwarepaket implementiert.

# 3 Material und Methoden

In diesem Kapitel werden zunächst die Monte-Carlo Simulationen einschließlich der zugrunde liegenden Annahmen und Modelle im Detail vorgestellt (Abschnitt 1). Hierzu wird zunächst das notwendige Modell zur Simulation des QTL und der Phänotypen nach Falconer und Mackay (1996) eingeführt und erläutert. Anschließend werden dann auch die im Rahmen dieser Arbeit verwendeten Familienstrukturen sowie Selektion der Familien und die konkreten simulierten Modelle definiert. Abschließend wird die Erstellung der simulierten Datensätze mit der Simulationssoftware `Sibsim` erläutert.

Der Abschnitt zwei gibt einen Überblick über die acht hier zu vergleichenden kopplungsanalytischen Verfahren, deren Softwareimplementationen sowie der konkret verwendeten Analyseparameter, die zur Analyse der Monte-Carlo Simulationen verwendet wurden. Anschließend wird dann die Berechnung der für den Gütevergleich erforderlichen empirischen Typ I-Fehler sowie der empirischen Power definiert.

Abschließend wird dann der COAG Perth Datensatz vorgestellt, sowie im letzten Abschnitt die verwendete Computerumgebung beschrieben.

## 3.1 Monte-Carlo Simulationen

### 3.1.1 QTL und Phänotypen

Als Basis für die Simulation der QTL und der dazugehörigen Phänotypen diente ein additives Modell mit einem biallelischen Mendelschen Hauptgen nach Falconer und Mackay (1996). Hierbei ist der Phänotyp $x_{im}$ der Person $i$ in der Familie $m$ additiv zerlegt in einem allgemeinen Mittelwert $\mu$, einen Hauptgeneffekt $g_{im}$, der durch den Genotyp eines biallelischen QTL bestimmt wird, einen Umwelteffekt $G_m$ simuliert als Familieneffekt und einen Fehlerterm $\varepsilon_{im}$:

$$x_{im} = \mu + g_{im} + G_m + \varepsilon_{im}$$

Es wird weiterhin angenommen, dass $g_{im}$, $G_m$ sowie $\varepsilon_{im}$ unkorreliert sind. Es ergibt sich dann als Varianz für den Phänotyp:

$$\sigma_x^2 = \sigma_g^2 + \sigma_G^2 + \sigma_\varepsilon^2$$

Der Hautgeneffekt wird dabei durch einen biallelischen Locus zusammen mit seinem spezifischen Vererbungsmodell bestimmt. Die Allele des Hauptgens seien dabei $A_1$ und $A_2$ mit den Frequenzen $p$ und $q = 1 - p$. Die Frequenz $p$ wird im Folgenden als Frequenz des hohen Allels bezeichnet.

Der Hauptgeneffekt wurde dann wie folgt modelliert:

$$g_{ik} = \begin{cases} a \text{ für eine Person mit Genotyp } A_1A_1 \\ d \text{ für eine Person mit Genotyp } A_1A_2 \\ -a \text{ für eine Person mit Genotyp } A_2A_2 \end{cases}$$

Im Rahmen dieser Arbeit wurden folgende Modelle betrachtet: Das dominante Modell (mit $a = d$), das additive Modell (mit $d = 0$) und das rezessive Modell (mit $d = -a$). Ein Modell mit Unterdominanz (mit $d < -a$) oder Überdominanz (mit $d > a$) wurde nicht betrachtet.

Die Erwartungswerte der Genotypen des Hauptgeneffektes wurden jedoch für die Simulation der Datensätze so verschoben, dass sich durch die Frequenzen $p$ und $q$ des biallelischen Locus und der zu simulierenden Varianz des Hauptgeneffektes ein Erwartungswert für den Hauptgeneffekt von $E(g_{ik}) = 0$ ergibt (siehe Ziegler und König, 2010, S. 160). Der Umwelteffekt wurde als Familieneffekt simuliert, bei dem jedem Mitglied der Familie $m$ der gleiche zufällige Wert zugeordnet wurde.

Um den Effekt der Verletzung der Normalverteilungsannahmen beurteilen zu können, wurde der Fehlerterm $\varepsilon_{im}$ des additiven Modells zum einen aus einer Normalverteilung und zum anderen aus einer logarithmischen Normalverteilung simuliert. Um eine logarithmische Normalverteilung mit gegebenen Mittelwert und Varianz zu erhalten, wurde zunächst die Standardnormalverteilung als Argument an *exp( )* verwendet, sodann der Mittelwert der zu simulierenden logarithmische Normalverteilung abgezogen und schließlich durch die Standardabweichung der zu simulierenden logarithmische Normalverteilung geteilt.

Insgesamt wurden also drei genetische Modelle und zwei Verteilungen betrachtet. Die Tabelle 2 und Tabelle 3 stellen eine Übersicht über die simulierten Modelle der Phänotypen dar.

**Tabelle 2:** Übersicht über die drei simulierten genetischen Modelle.

| Genetisches Modell | Frequenz „hohes Allel“ | $\mu$ | $\delta_g^2$ | $\delta_G^2$ | $\delta_\varepsilon^2$ | $a$ | $d$ |
|---|---|---|---|---|---|---|---|
| Dominant | 0,05 | 0 | 0,2 | 0,3 | 0,5 | 0,754 | 0,754 |
| Additiv | 0,2 | 0 | 0,2 | 0,3 | 0,5 | 0,474 | 0 |
| Rezessiv | 0,3 | 0 | 0,2 | 0,3 | 0,5 | 0,782 | -0,782 |

Wie aus der Tabelle 2 ersichtlich, gilt für alle Modelle der Phänotypen $\mu = 0$, eine Varianz von 1 und eine Heritabilität von 0,2 im weiteren Sinne und von 0,5 im engeren Sinne.

**Tabelle 3:** Erwartungswerte des Hauptgeneffektes für die drei verwendeten genetischen Modelle.

| Genetisches Modell | $E(A_1,A_1)$ | $E(A_1,A_2)$ | $E(A_2,A_2)$ |
|---|---|---|---|
| Dominant | 1,361 | 1,361 | -0,147 |
| Additiv | 1,265 | 0,474 | -0,316 |
| Rezessiv | 1,422 | -0,141 | -0,141 |

### 3.1.2 Familienstrukturen

Die simulierten Datensätze wurden auf zwei verschiedenen Familienstrukturen basierend simuliert. Als Ausgangspunkt wurde dabei die Struktur einer Kernfamilie mit einem Geschwisterpaar gewählt. Dieser Begriff meint einen Stammbaum mit der Struktur eines Elternpaars als Gründer in der ersten Generation und eines Geschwisterpaars als Abkömmlinge. Da es nur ein Geschwisterpaar pro Familie gibt, sind die Geschwisterpaare unabhängig. Als zweite Familienstruktur sollte eine Kernfamilie mit abhängigen Geschwisterschaften betrachtet werden, d.h. eine Mischung aus Kernfamilien mit einer unterschiedlichen Anzahl von Geschwistern als Abkömmlingen. Die Mischung sollte hierbei einer realistischen Verteilung in Industrieländern entsprechen. Hier wurden deshalb die Verteilung der Familienstrukturen vom simulierten Datensatz für den Genetic Analyses Workshop 9 verwendet (Speer et al., 1995). Diese Familienstruktur wird im Folgenden als Kernfamilie mit abhängigen Geschwisterschaften bezeichnet. Die genaue Zusammensetzung ergibt sich Tabelle 4.

**Tabelle 4**: Verteilung der Geschwisterschaften unter der Familienstruktur abhängigen Geschwisterschaften.

| Anzahl Geschwister | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Anteil | 40 % | 30 % | 18 % | 12 % |

Pro Datensatz wurden 300 Familien unter der Familienstruktur Kernfamilie mit einem Geschwisterpaar und 100 Familien unter der Familienstruktur Kernfamilie mit mehrfachen Geschwisterschaften simuliert. Die Phänotypen und Genotypen wurden für alle Familienmitglieder simuliert.

### 3.1.3 Genetischer Marker

Für jeden Datensatz innerhalb einer Simulation wurde ein genetischer Marker mit zehn Allelen gleicher Häufigkeit simuliert. Unter der Nullhypothese wurde der genetische Marker mit einem Abstand von $\theta = 0{,}5$ zum QTL und unter der Alternativhypothese mit einem Abstand von $\theta = 0$ zum QTL simuliert.

### 3.1.4 Studiendesign: Selektion von Familien

Um den Effekt der Selektion von Familien beurteilen zu können wurden drei verschiedene Studiendesigns verwendet: Es wurden zum einen Familien unter zufälliger Selektion, Familien mit wenigstens einem Kind im oberen Quartil der Phänotypverteilung (einfache Selektion, entspricht einem SPSP-Design) und Familien mit entweder zwei Kinder im oberen Quartil, zwei Kinder im unteren Quartil oder einem Kind im oberen Quartil und einem Kind im unteren Quartil der Phänotypverteilung (doppelte Selektion, entspricht einem ESP Design) gewählt. Die Simulation der Datensätze wurde unter Selektion so lange fortgesetzt, bis 100.000 Datensätze unter der Nullhypothese und 1.000 Datensätze unter der Alternativhypothese simuliert wurden, welche die Selektionskriterien erfüllen.

### 3.1.5 Simulation der Datensätze

Zur Erstellung der simulierten Datensätze wurde das am IMBS erstellte Programm `Sibsim` Version 1.02 verwendet (Franke et al., 2006). Die Datensätze wurden im `Linkage`-Datenformat erstellt und bei Bedarf für die verwendeten Analyseprogramme mit der Software `Mega2` in das jeweils notwendige Datenformat konvertiert (Mukhopadhyay et al., 2005). Für die Definition des `Linkage`-Datenformats siehe z.B. Terwilliger und Ott (1994, Kapitel 2).

Eine Simulation bestand jeweils aus 100.000 Datensätzen unter der Nullhypothese ($\theta = 0{,}5$) und 1.000 Datensätzen unter der Alternativhypothese ($\theta = 0$).

Insgesamt wurden also 36 verschiedene Modelle (3 genetische Modelle × 2 Verteilungen × 2 Familienstrukturen × 3 Selektionsschemata) simuliert wie zuvor im Detail beschrieben.

Die Startzufallszahlen für das Softwarepaket `Sibsim` sind in Tabelle 23 im Anhang dokumentiert.

## 3.2 Gütevergleich

### 3.2.1 Angewandte kopplungsanalytische Verfahren

Zur Analyse der Simulationen wurden insgesamt acht verschiedene kopplungsanalytische Verfahren implementiert in sechs Softwarepaketen verwendet. Alle verwendeten Softwarepakete waren beim Erscheinen dieser Arbeit zur nichtkommerziellen Nutzung frei verfügbar. Die verwendeten Versionen und Quellen der Softwarepakete sind im Anhang Kapitel 7, Abschnitt 1 zu finden. In der folgenden Tabelle sind die verwendeten Softwarepakete den verwendeten kopplungsanalytischen Verfahren gegenübergestellt. Das HASEMAN-ELSTON Verfahren sowie die VARIANZKOMPONENTENMODELLE wurden zum Vergleich der Implementierung mit zwei verschiedenen Programmen berechnet.

**Tabelle 5**: Übersicht über die verwendeten kopplungsanalytischen Verfahren, verwendete Abkürzungen und die dazugehörigen Softwarepakete.

| Programm | Kopplungsanalytisches Verfahren | Verwendete Abkürzungen |
|---|---|---|
| `Genehunter` | HASEMAN-ELSTON Verfahren | Gh.HE.Trad |
| | WILCOXON-RANGSUMMENTEST | Npar |
| | VARIANZKOMPONENTENMODELLE | Gh.VC |
| `Linkage` | LINKAGE (MODELLBASIERTES Verfahren) | Linkage |
| `Merlin` | MERLIN-REGRESS | Merlin-Regress |
| | MERLIN-QTL | Merlin.K&C, Merlin.W&H |
| `Mlbgh` | MAXIMUM LIKELIHOOD BINOMIAL Verfahren für quantitative Phänotypen | MLBQT.NV, MLBQT.Kat |
| `S.A.G.E.` | HASEMAN-ELSTON Verfahren | Sage.HE |
| | REVIDIERTES HASEMAN-ELSTON Verfahren | rHE |
| `Solar` | VARIANZKOMPONENTENMODELLE | Solar.VC |

Im Folgenden werden jeweils die verwendeten Programme sowie die verwendeten Analyseparameter in alphabetischer Reihenfolge erläutert.

Genehunter

Mit dem Programm Genehunter wurden das HASEMAN-ELSTON Verfahren (Gh.HE.Trad), der WILCOXON-RANGSUMMENTEST (Npar) sowie die VARIANZKOMPONENTENMODELLE (Gh.VC) gerechnet. Für alle Analysen wurde die Option alle Geschwister ungewichtet verwendet („all pairs unweighted"). Die VARIANZKOMPONENTENMODELLE wurden ohne Dominanzterme berechnet. Da die elterlichen Genotypen bekannt sind, wurden beim HASEMAN-ELSTON Verfahren die Berechnungen ohne Anwendung des EM-Algorithmus durchgeführt (Dempster et al., 1977).

Linkage

Mit dem Softwarepaket Linkage wurde das MODELLBASIERTE Verfahren (Linkage) berechnet. Für die Analysen wurden jeweils die Erwartungswerte des Hauptgeneffektes, die dazugehörigen Frequenzen für den biallelischen QTL sowie die populationsbasierten Varianzen der Verteilung der Phänotypen verwendet. Die Erwartungswerte der drei genetischen Modelle sind in Tabelle 3 (S. 19), die Allelfrequenzen des biallelischen QTL in Tabelle 2 (S. 19) gegeben.

Merlin

Mit dem Programm Merlin-regress, welches Bestandteil des Merlin Softwarepaketes ist, wurde das Verfahren von Sham et al. (2002) berechnet (Merlin-regress). Als Eingabeparameter wurden die wahren populationsbasierten Werte für Mittelwert ($\mu = 0$), Varianz ($\sigma^2 = 1$) und Heritabilität im engeren Sinne ($h^2 = 0{,}5$) übergeben. Um einen möglichen Einfluss von Model-Missspezifikationen zu ermitteln, wurden alle Datensätze zusätzlich unter einer Vielzahl von ein Parameter Missspezifikationen für Mittelwert (-5 bis 5), Varianz (0,1 bis 10) und Heritabilität (0,05 bis 0,95) analysiert.

Das Verfahren MERLIN-QTL wurde mit dem Programm Merlin aus dem Merlin-Softwarepaket berechnet. Es wurden sowohl der Ansatz mit der Teststatistik nach Whittemore und Halpern (1994), als auch der LQT nach Kong und Cox (1997) angewendet.

Mlbgh

Bei dem Programm Mlbgh handelt es sich um eine Modifikation des Programms Genehunter. Es wurde für das MAXIMUM LIKELIHOOD BINOMIAL Verfahren verwendet.

Die Analysen wurden sowohl unter der Annahme einer Standardnormalverteilung der Phänotypen (Mlbqt.NV), als auch unter Verwendung einer Treppenfunktion zur Definition der Schwellenwerte

durchgeführt (Mlbqt.Kat). Die Schwellenwerte wurden dabei durch Einteilung der Verteilung der Phänotypen in empirische Dezile definiert, wie von Alcaïs und Abel (1999) vorgeschlagen.

S.A.G.E.

Das Programmpaket S.A.G.E. wurde verwendet um das HASEMAN-ELSTON (Sage.HE) und das REVIDIERTE HASEMAN-ELSTON Verfahren (Sage.rHE) anzuwenden. Die Analyse wurde unter den vorgegebenen Standardoptionen durchgeführt. Für das REVIDIERTE HASEMAN-ELSTON Verfahren wurde der wahre populationsbasierte Wert für den Mittelwert ($\mu = 0$) verwendet.

Solar

Neben dem Genehunter Softwarepaket wurde Solar verwendet, um die VARIANZ-KOMPONENTENMODELLE zu berechnen (Solar.VC). Es wurden die Standardparameter verwendet.

### 3.2.2 Empirischer Typ I-Fehler und empirische Power

Der empirische Typ I-Fehler wurde auf Basis der 100.000 Simulationen unter der Nullhypothese bei einem nominalen Typ I-Fehlerniveau von 5%, 1% und 0,1% errechnet. Standardnormalverteilte *Z*-Scores oder LOD-Scores wurden in *p*-Werte konvertiert. Zur Konvertierung der LOD-Scores siehe z.B. Ziegler und König (2010, S. 159).
Zusätzlich wurde der Grad der Abweichung zwischen den nominalen und empirischen Typ I-Fehlern durch Einteilung in sieben Kategorien bei einem nominalen Typ I-Fehlerniveau von 1% berechnet. Der Normalbereich (○) wurde hierbei durch ein 95% Clopper-Pearson Konfidenzintervall aus einer Binomialverteilung mit 100,000 Versuchen und 1,000 Erfolgen Bonferroni korrigiert für 36 Modelle ermittelt, während die weiteren Grenzen willkürlich bei einer Abweichung von Faktor 1,333 und 2 bezogen auf den nominalen Typ I-Fehler festgelegt wurden (siehe Abbildung 1).

⇩⇩⇩ 0,5% ⇩⇩ 0,750% ⇩ 0,902% ○ 1,105% ⇧ 1,333% ⇧⇧ 2,0% ⇧⇧⇧

**Abbildung 1**: Empirischer Typ I-Fehler bei einem nominalen Typ I-Fehler von 1% eingeteilt in sieben Kategorien. Mit ⇩ ist ein konservativen und mit ⇧ ein liberaler Typ I-Fehler bezeichnet.

Die empirische Power wurde als die Wahrscheinlichkeit berechnet, dass die Teststatistik unter der Alternativhypothese die aus der empirischen Verteilung unter der Nullhypothese errechneten Signifikanzschwelle bei einem gegebenen empirischen Typ I-Fehler von 0,05 überschreitet.

## 3.3 COAG Perth Datensatz

Der Datensatz vom „Consortium on Asthma Genetics: Perth study“ (COAG Perth Datensatz), wie er für den Genetic Analysis Workshop 12 zur Verfügung gestellt worden ist, wurde als praktisches Beispiel für eine Kopplungsanalyse mit einem quantitativem Merkmal ausgewählt (Für eine detaillierte Beschreibung siehe: Palmer et al., 1998; Palmer et al., 2001). Die Verwendung dieser Daten erfolgte mit freundlicher Genehmigung von Herrn Dr. Palmer vom Department of Pediatrics and Centre for Molecular Immunology and Instrumentation, University of Western Australia, Perth, Australien.

Der Original-Datensatz bestand aus 123 Familien mit insgesamt 583 Personen, die im Raum Perth (Australien) rekrutiert wurden. Hierbei wurden 25 Kernfamilien spezifisch im Hinblick auf Asthma-Erkrankungen rekrutiert, während 98 Kernfamilien zufällig ausgewählt wurden. Für die erneute Analyse wurde auf die zufällig ausgewählten 98 Familien zurückgegriffen. Für eine Kandidatenregion auf Chromosom 5q31-33 standen zwei gekoppelte typisierte Mikrosatelliten-Marker (D5S393 und D5S399) mit der Rekombinationsfrequenz von $\theta = 0{,}0006$ zur Verfügung.

Von den im Datensatz zur Verfügung stehenden asthmarelevanten Phänotypen wurden der logarithmierte Gesamt Serum IgE Titer (ln IgE) sowie der logarithmierte Gesamt Serum IgE Titer adjustiert für Alter, Geschlecht und antigen-spezifischen IgE Titer (ln IgER) für die Kopplungsanalysen verwendet.

## 3.4 Verwendete Computerumgebung

Alle durchgeführten Berechnungen und Simulationen wurden auf Intel© Dual Xeon© Prozessoren mit 2,8 GHz Taktfrequenz unter dem Betriebssystem Suse Linux Version 8.1 bzw. Version 8.2 (Linux Kernel Version 2.4.21) durchgeführt. Die verwendeten Softwarepakete wurden, soweit verfügbar, in einer für Linux kompilierten Version direkt verwendet. Wenn eine vorkompilierte Version nicht verfügbar war, wurden die Programme selbst kompiliert. Alle verwendeten Programme wurden mit den jeweils beiliegenden Testdatensätzen auf Korrektheit der Berechnungen in der verwendeten Computerumgebung überprüft. Eine Liste der verwendeten Softwarepakete befindet sich in Kapitel 7, Abschnitt 1.

# 4 Ergebnisse

Das Kapitel Ergebnisse gliedert sich in fünf Abschnitte. Zunächst wird im ersten Abschnitt die systematische externe Validierung der Simulationssoftware `Sibsim` dargestellt. Dieser Abschnitt gliedert sich in eine Übersicht, in der die Zielkriterien der externen Validierung definiert werden; darauf folgt dann die Darstellung der Ergebnisse der einzelnen Validierungen.
Abschnitt 2 gibt einen kurzen Überblick über die simulierten Datensätze sowie die Berechnung der Teststatistiken.
Die Abschnitte 3 und 4 stellen sodann den Gütevergleich der kopplungsanalytischen Verfahren dar. In Abschnitt 3 werden die Ergebnisse des Vergleiches der empirischen Typ I-Fehler zu den nominalen Typ I-Fehlern für jedes einzelne Verfahren separat dargestellt und erläutert. Die Ergebnisse des empirischen Powervergleichs sind im Abschnitt 4 dargestellt. Im ersten Unterabschnitt wird zunächst die empirische Power innerhalb der Verfahren unter den verschieden Szenarien miteinander verglichen. Im zweiten Unterabschnitt werden dann innerhalb eines Szenarios die Verfahren direkt miteinander verglichen. Der dritte Unterabschnitt fasst die Ergebnisse dann noch einmal abhängig von der jeweiligen Verfahrensweise zusammen.
Abschließend wird in Abschnitt 5 die Analyse des COAG Perth Datensatzes detailliert gezeigt. Besonderer Wert wurde hierbei darauf gelegt, die praktische Durchführung einer genetischen Kartierung quantitativer Merkmale zu illustrieren.

## 4.1 Externe Validierung der Simulationssoftware Sibsim

Im Rahmen dieser Arbeit wurde das Softwarepaket `Sibsim` (Franke et al., 2006) von Herrn Dr. Franke und dem Verfasser erstellt. Vor der Verwendung für die notwendigen Monte-Carlo Simulationen hat der Verfasser `Sibsim` dann mit einer systematischen externen Validierungsprozedur überprüft. Die systematische externe Validierung wird hier als externe dokumentierte Überprüfung für den speziell beabsichtigten Gebrauch im Rahmen dieser Arbeit verstanden. Hierbei sollte sichergestellt werden, dass die mit `Sibsim` erstellten simulierten Datensätze den geforderten Simulationsparametern entsprechen.
Das Softwarepaket `Sibsim` wurde deshalb nach den folgenden Parametern validiert:

- Entspricht die Ausgabe von `Sibsim` in Hinsicht auf Familienstrukturen, -größen und Anzahl der Familien pro Datensatz den Simulationsparametern?
- Entsprechen die Allelfrequenzen and Anzahl der Allele für die QTL- sowie Marker-Genotypen den Simulationsparametern?
- Entsprechen die Genotypen des QTL- sowie der Marker-Genotypen der Kinder den Mendelschen-Vererbungsregeln?

- Entsprechen die simulierten Phänotypen den Simulationsparametern in Hinblick auf
    - -- Gesamt-Mittelwert und -Varianz,
    - -- Mittelwert und Varianz der Verteilung des Hauptgeneffektes, Fehlerterms und Familieneffektes, sowie
    - -- der Verteilungsform?
- Entsprechen die simulierten Datensätze unter Selektion den Selektionskriterien?

Hierzu wurden jeweils für das dominante, additive und rezessive genetische Modell unter geeigneten Simulationsparametern mit der Familienstruktur abhängige Geschwisterschaften eine Simulation mit zehn Datensätzen erstellt. Insgesamt wurden acht Simulationen erstellt, die im Folgenden als Validierungs-Simulationen bezeichnet werden. Die verwendeten Simulationsparameter sind im Anhang Tabelle 22, S. 66 dargestellt. Die Validierungs-Simulationen 1, 2 und 3 basieren auf einem dominanten, rezessiven und additiven genetischen Modell das einen Hauptgeneffekt von 2 sowie einen Fehlerterm von 0,2 enthielt. Ein Familieneffekt wurde nicht simuliert.
Die Validierungs-Simulationen 4 und 5 basieren auf einem dominanten genetischen Modell, das einen Hauptgeneffekt von 4 sowie einen Fehlerterm von 0,5 enthielt. In der Simulation 5 wurde der Fehlerterm aus einer logarithmierten Normalverteilung simuliert. Die Validierungs-Simulationen 6, 7 und 8 basieren auf einem dominanten, rezessiven und additiven genetischen Modell das einen Hauptgeneffekt von 2 sowie einen Familieneffekt von 0,2 enthielt. Ein Fehlerterm wurde nicht simuliert. Aus diesen zehn Datensätzen wurde dann per Zufall ein Datensatz pro genetisches Modell zur Validierung ausgewählt.
Zusammenfassend lässt sich sagen, dass die systematische externe Validierung keine Auffälligkeiten ergab.

### 4.1.1 Familienstrukturen, -größen und Anzahl der Familien pro Datensatz

In allen Datensätzen der Validierungs-Simulationen eins bis acht sowie in allen im Rahmen dieser Arbeit zum Gütevergleich verwendeten Datensätzen wurden die Familienstrukturen, -größen sowie die Anzahl der Familien pro Datensatz überprüft. Die Ergebnisse stimmen mit den geforderten Simulationsparametern überein (Ergebnisse hier nicht im Detail gezeigt).

### 4.1.2 Vererbungsregeln und Allelfrequenzen für die QTL- sowie Marker-Genotypen

Im ersten Schritt wurden zunächst die Allelfrequenzen für die QTL- sowie Markergenotypen geschätzt sowie die Mendelschen-Vererbungsregeln mit dem Softwarepaket `Pedcheck` überprüft. Um die Allelfrequenzen und Vererbungsregeln der QTL-Genotypen zu überprüfen, wurde das Softwarepaket `Sibsim` hierfür so modifiziert, dass es zusätzlich die Allele des biallelischen QTL

ausgibt. Aus den Validierungs-Simulationen 1, 2 und 3 wurden hierfür die Datensätze 10, 7 und 6 zufällig ausgewählt. Die Ergebnisse der geschätzten Allelfrequenzen für die Marker- sowie QTL-Genotypen zeigen nur geringe, im Rahmen der zufälligen Streuung auftretende Abweichungen im Vergleich zu den geforderten Simulationsparametern (Ergebnisse hier nicht im Detail gezeigt). Es wurden keine Verletzungen der Mendelschen-Vererbungsregeln gefunden.

### 4.1.3 Simulation der Phänotypen

#### 4.1.3.1 Hauptgeneffekt sowie Fehlerterm

Basierend auf den Validierungs-Simulationen 1, 2 und 3 wurde zufällig jeweils einer der 10 simulierten Datensätze zur Validierung des Hauptgeneffektes sowie des Fehlerterms ausgewählt. Dieses waren aus den Validierungs-Simulationen 1, 2 und 3 die Datensätze 6, 7 und 4. Die Verteilung der Phänotypen ist in Abbildung 2 dargestellt.

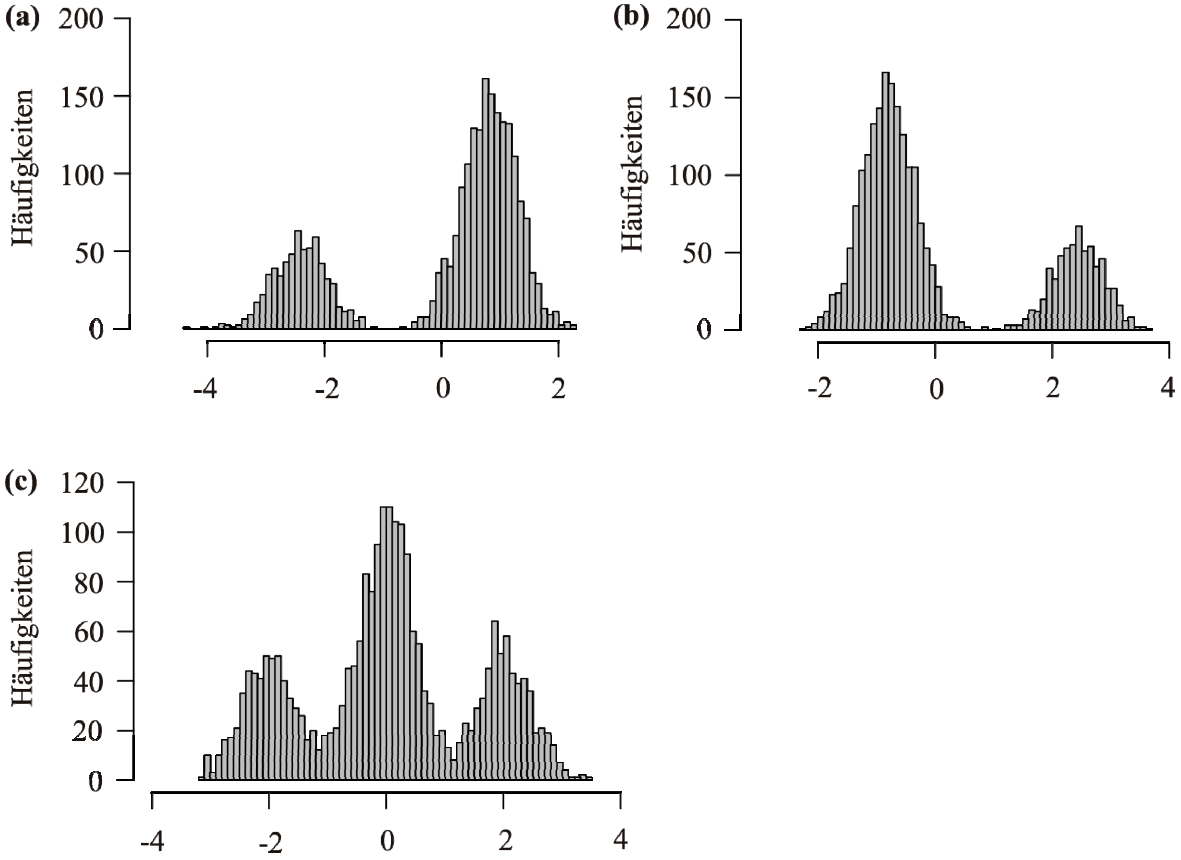


**Abbildung 2**: Histogramm der Verteilungen der Phänotypen für ausgewählte Validierungs-Simulationen unter einem dominanten (a, Simulation 1, Datensatz 6), rezessiven (b, Simulation 2, Datensatz 7) und einem additivem (c, Simulation 3, Datensatz 4) Vererbungsmodell mit Hauptgeneffekt und Fehlerterm.

Basierend auf den Erwartungswerten der zwei bzw. drei Hauptgeneffekten wurden Grenzen definiert, die die zwei bzw. drei Verteilungen voneinander abgrenzen sollen (siehe Tabelle 6). Da zusätzlich zum Hauptgeneffekt auch ein Fehlerterm simuliert wurde, kommt es zu geringen

Überschneidungen zwischen den Verteilungen. Die Mittelwerte und Varianzen des gesamten Phänotyps sowie die Mittelwerte und Häufigkeiten für den Hauptgeneffekt und die Varianzen der Fehlerterme sind in Tabelle 6 dargestellt. Es sind nur kleine durch den Zufall zu erklärende Differenzen zu erkennen.

**Tabelle 6:** Beobachtete sowie erwartete Mittelwerte, Varianzen und Häufigkeiten für die Validierungs-Simulationen 1, 2 und 3 mit den zufällig ausgewählten Datensätzen 6, 7 und 4 sowie deren Hauptgeneffekte für ein Modell mit Hauptgeneffekt und Fehlerterm.

| | Dominant | | | Additiv | | | | Rezessiv | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | < -0,817 | ≥ 0,817 | Gesamt | < -1 | ≥ -1 und ≤ 1 | > 1 | Gesamt | < 0,817 | ≥ 0,817 | Gesamt |
| Beobachtet: | | | | | | | | | | |
| Mittelwert | -2,477 | 0,824 | 0,025 | -1,994 | -0,001 | 1,993 | -0,105 | -0,840 | 2,413 | 0,001 |
| Varianz | 0,191 | 0,198 | 2,196 | 0,200 | 0,193 | 0,193 | 2,295 | 0,209 | 0,188 | 2,236 |
| Anzahl | 581 | 1819 | 2400 | 703 | 1120 | 577 | 2400 | 1779 | 621 | 2400 |
| Erwartet: | | | | | | | | | | |
| Mittelwert | -2,45 | 0,816 | 0 | -2 | 0 | 2 | 0 | -0,817 | 2,45 | 0 |
| Varianz | 0,2 | 0,2 | 2,2 | 0,2 | 0,2 | 0,2 | 2,2 | 0,2 | 0,2 | 2,2 |
| Anzahl | 600 | 1800 | 2400 | 600 | 1200 | 600 | 2400 | 1800 | 600 | 2400 |

Die simulierte Normalverteilung bzw. logarithmierte Normalverteilung des Fehlerterms wurde mit einem Normalverteilungsplot bzw. einem logarithmierten Normalverteilungsplot überprüft. Da die logarithmierte Normalverteilung aber rechts schief ist, wurde der Hauptgeneffekt im Vergleich zu den Validierungs-Simulationen zuvor erhöht, um die Verteilungen besser von einander trennen zu können (Validerungs-Simulation 5). Hierzu wurden für das dominante genetische Modell aus den Validierungs-Simulationen 4 und 5 die Datensätze 9 und 3 zufällig ausgewählt.

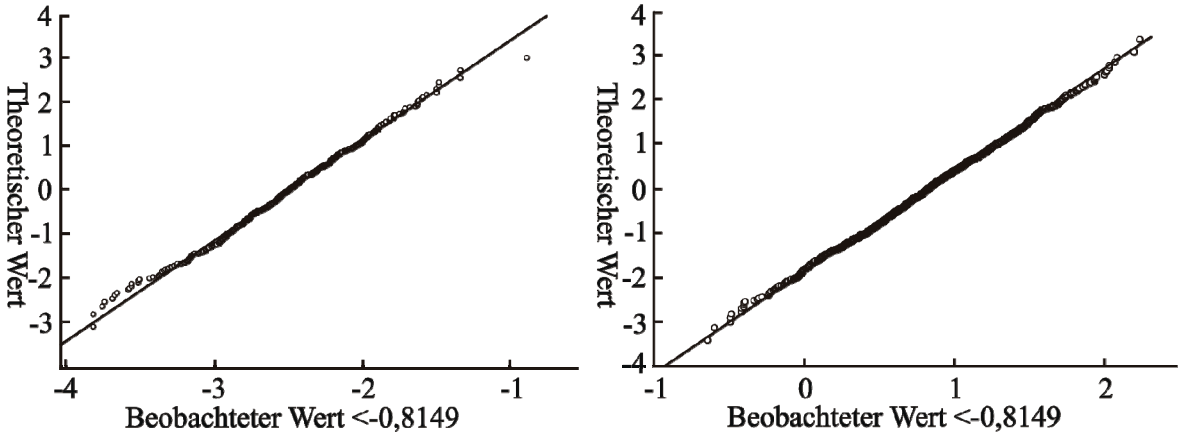


**Abbildung 3**: Normalverteilungsplot für Validierungs-Simulation 4, Datensatz 9 zur Überprüfung der Normalverteilung des Fehlerterms.

Wie in Abbildung 3 und in Abbildung 4 zu erkennen ist, folgen die Verteilungen des Fehlerterms einer Normal- bzw. logarithmierten Normalverteilung. Die Ergebnisse für das additive und rezessive Modell sind vergleichbar, jedoch hier nicht extra gezeigt.

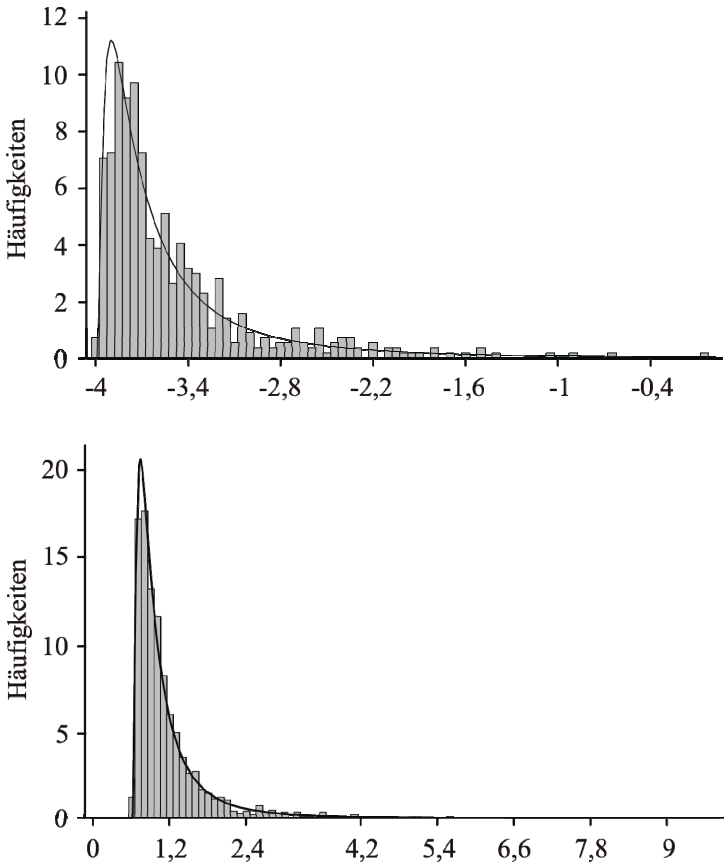


**Abbildung 4**: Logarithmierte Normalverteilungsplots für Validierungs-Simulation 5 Datensatz 3 zur Überprüfung der logarithmierten Normalverteilung des Fehlerterms.

Die Mittelwerte und Varianzen des gesamten Phänotyps sowie die Mittelwerte und Häufigkeiten für die Hauptgeneffekte und die Varianzen der Fehlerterme sind in Tabelle 7 dargestellt. Wie zuvor sind nur kleine durch den Zufall zu erklärende Differenzen zu erkennen.

**Tabelle 7**: Beobachtete sowie erwartete Mittelwerte, Varianzen und Häufigkeiten für die Validierungs-Simulationen 4, 5 mit den zufällig ausgewählten Datensätzen 9 und 3 sowie deren Hauptgeneffekte für ein Modell mit Hauptgeneffekt und Fehlerterm.

| | Normalverteilung Fehlerterm | | | Logarithmierte Normalverteilung Fehlerterm | | |
|---|---|---|---|---|---|---|
| | $< -1{,}155$ | $\geq -1{,}155$ | Gesamt | $< 0$ | $\geq 0$ | Gesamt |
| Beobachtet: | | | | | | |
| Mittelwert | -3,439 | 1,122 | -0,091 | -3.495 | 1.152 | 0.017 |
| Varianz | 0,532 | 0,469 | 4,533 | 0.313 | 0.511 | 4.451 |
| Anzahl | 635 | 1757 | 2400 | 586 | 1814 | 2400 |
| Erwartet: | | | | | | |
| Mittelwert | -3,464 | 1,155 | 0 | -3,464 | 1,155 | 0 |
| Varianz | 0,5 | 0,5 | 4,5 | 0,5 | 0,5 | 4,5 |
| Anzahl | 600 | 1800 | 2400 | 600 | 1800 | 2400 |

### 4.1.3.2 Familieneffekt

Zur Überprüfung des simulierten Familien Effektes wurden aus den Validierungs-Simulationen 6, 7 und 8 die Datensätze 10, 7 und 8 zufällig ausgewählt, in denen nur ein Hauptgeneffekt sowie ein Familieneffekt simuliert wurden. Die Mittelwerte und Varianzen des gesamten Phänotyps sowie die Mittelwerte und Häufigkeiten für den Hauptgeneffekt und die Varianzen der Familieneffekte sind in Tabelle 8 dargestellt. Wiederum sind nur kleine durch den Zufall zu erklärende Differenzen zu erkennen.

**Tabelle 8**: Beobachtete sowie erwartete Mittelwerte, Varianzen und Häufigkeiten für die Validierungs-Simulationen 6, 7 und 8 mit den zufällig ausgewählten Datensätzen 10, 7 und 8 sowie deren Hauptgeneffekte für ein Modell mit Hauptgen- und Familieneffekt.

| | Dominant | | | Additiv | | | | Rezessiv | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | < -0,817 | ≥ 0,817 | Gesamt | < -1 | ≥ -1 und ≤ 1 | > 1 | Gesamt | < 0,817 | > 0,817 | Gesamt |
| Beobachtet: | | | | | | | | | | |
| Mittelwert | -2,446 | 0,801 | 0,013 | -1,973 | 0,008 | 2,0118 | 0,007 | -0,814 | 2,424 | 0,002 |
| Varianz | 0,205 | 0,209 | 2,145 | 0,191 | 0,181 | 0,177 | 2,227 | 0,200 | 0,226 | 2,184 |
| Anzahl | 582 | 1818 | 2400 | 622 | 1164 | 614 | 2400 | 1795 | 605 | 2400 |
| Erwartet | | | | | | | | | | |
| Mittelwert | -2,45 | 0,816 | 0 | -2 | 0 | 2 | 0 | -0,817 | 2,45 | 0 |
| Varianz | 0,2 | 0,2 | 2,2 | 0,2 | 0,2 | 0,2 | 2,2 | 0,2 | 0,2 | 2,2 |
| Anzahl | 600 | 1800 | 2400 | 600 | 1200 | 600 | 2400 | 1800 | 600 | 2400 |

Im Gegensatz zum Fehlerterm ist der Familieneffekt für alle Mitglieder einer Familie gleich. Da die Validierungs-Simulationen 6, 7 und 8 keinen Fehlerterm enthalten, ergibt sich daraus, dass es nur max. zwei für das dominante bzw. drei für das rezessive Modell mögliche Phänotypen innerhalb einer Familie geben kann. Dieses wurde in den drei Datensätzen ebenfalls überprüft.

## 4.1.4 Selektion der Familien

Zur Überprüfung der Selektion von Familien wurden die simulierten Datensätze zum Vergleich der kopplungsanalytischen Verfahren unter Normalverteilungsannahmen und unabhängige Geschwisterschaften verwendet. Zunächst wurden auf der Basis von 30.000 simulierten Phänotypen für die drei Modelle unter zufälliger Selektion die empirischen Quartilsgrenzen geschätzt. Im nächsten Schritt wurde für einen zufällig ausgewählten Datensatz unter einfacher Selektion das Maximum der Phänotypen der Kinder pro Familie ermittelt. Das Minimum der ermittelten Familienmaxima sollte dann in etwa der zuvor ermittelten empirischen oberen Quartilsgrenze entsprechen. Hierdurch wird zum einen validiert, dass alle Familien das einfache Selektionskrite-

rium erfüllen, und zum anderen, dass die Familien nicht auf einem strengeren Niveau selektiert wurden.

Die Ergebnisse sind in Tabelle 9 zusammengefasst. Die Ergebnisse zeigen die bei den gegebenen Stichprobengrößen von 100 Familien pro ausgewähltem Datensatz zu erwartenden Differenzen.

**Tabelle 9**: Validierung der einfachen Selektion für zufällig ausgewählte Datensätze unter einem dominanten, additiven und rezessiven Modell. Geschätzte obere und untere Quartilsgrenzen unter zufälliger Selektion sowie Minimum der Familienmaxima unter einfacher Selektion, jeweils geschätzt in den Kindern.

| | Dominant | Additiv | Rezessiv |
|---|---|---|---|
| Geschätzte obere Quartilsgrenze | 0,635 | 0,667 | 0,621 |
| Geschätzte untere Quartilsgrenze | -0,692 | -0,676 | -0,673 |
| Ausgewählter Datensatz | 46803 | 21050 | 4933 |
| Minimum der Familienmaxima | 0,631 | 0,674 | 0,649 |

Zur Überprüfung der doppelten Selektion wurde jeweils für einen zufällig ausgewählten Datensatz unter doppelter Selektion für jede Familie ermittelt, ob sie einer der drei Möglichkeiten der doppelten Selektion entsprechen. Zusätzlich wurden das Minimum der über die obere Quartilsgrenze selektierten Kinder und das Maximum der über die untere Quartilsgrenze selektierten Kinder ermittelt.

**Tabelle 10:** Validierung der doppelten Selektion für zufällig ausgewählte Datensätze unter einem dominanten, additiven und rezessiven Modell. Es sind die Anzahl der Familien, die einem der drei möglichen Selektionskriterien entsprechen sowie das Minimum/Maximum der über das obere/untere Quartilsgrenze selektierten Kinder gezeigt. Mit Q4 wird dabei die obere Quartilsgrenze während mit Q1 die untere Quartilsgrenze bezeichnet wird.

| | Dominant | Additiv | Rezessiv |
|---|---|---|---|
| Ausgewählter Datensatz | 4989 | 28579 | 40900 |
| Selektierte Familien: | | | |
| 'Q4, Q4 | 42 | 43 | 53 |
| 'Q4, Q1 | 25 | 24 | 23 |
| 'Q1, Q1 | 46 | 41 | 31 |
| Familien die mind. eines der Kriterien erfüllen | 100 | 99 | 100 |
| Minimum der über die obere Quartilsgrenze selektierten Kinder | 0,636 | 0,675 | 0,624 |
| Maximum der über die unteren Quartilsgrenze selektierten Kinder | -0,693 | -0,689 | -0,678 |

Die Ergebnisse in Tabelle 9 und Tabelle 10 zeigen die bei den gegebenen Stichprobengrößen von 100 Familien pro ausgewählten Datensatz zu erwartenden zufälligen Differenzen.

## 4.2 Datensätze und Berechnung der Teststatistiken

Die Größe der simulierten Datensätze betrug insgesamt 130 GB. Die Familienstrukturen, -größen und Anzahl der Familien pro Datensatz wurden für alle Datensätze überprüft und entsprechen den

geforderten Simulationsparametern (hier nicht näher gezeigt). Zusätzlich wurden die Vererbungsregeln des genetischen Markers mit dem Softwarepaket `Pedcheck` für alle Datensätze überprüft.
Die Berechnung aller Teststatistiken benötigte ca. zwei Prozessorjahre auf einem Intel© Xeon© Prozessor mit 2,8 GHz Taktfrequenz, wobei die Berechnungen parallel auf mehreren Prozessoren durchgeführt wurden. Die Berechnungen der Teststatistiken wurden dabei über Bash-Skripte automatisch ausgeführt. Ebenso wurde die Berechnung der empirischen Typ I-Fehler sowie der empirischen Power über Bash-Skripte und `R`-Programme automatisiert durchgeführt. Die Teststatistiken sowie weitere Zwischenergebnisse ergaben weitere 4 GB an Daten.

## 4.3 Empirische Typ I-Fehler der Verfahren

Die Darstellung der empirischen Typ I-Fehler der Verfahren geht jeweils von einer tabellarischen Übersicht zum empirischen Typ I-Fehler des jeweiligen Verfahrens bei einem nominalen Typ I-Fehler von 0,01 aus. Dabei wird zusätzlich eine Einteilung in sieben Kategorien – wie in Kapitel 3, Abschnitt 2.2 beschrieben – vorgenommen (Tabelle 11 bis Tabelle 18). Ergänzend werden auch die Ergebnisse zum empirischen Typ I-Fehler bei einem nominalen Typ I-Fehler von 0,05 und 0,001 im Anhang in der Tabelle 24, S. 67 bis Tabelle 29, S. 72 dargestellt.

Unter Normalverteilungsannahmen, unabhängiger Geschwisterschaften und zufälliger Selektion zeigen die VARIANZKOMPONENTENMODELLE einen deutlich zu liberalen Typ I-Fehler während das MERLIN-QTL Verfahren mit der Whittemore und Halpern Teststatistik und das MODELBASIERTE Verfahren fast immer einen konservativen Typ I-Fehler zeigt. Die anderen Verfahren halten das korrekte Typ I-Fehlerniveau. Eine Abweichung bei einer oder mehrer dieser drei Annahmen hat vielfach einen Einfluss auf den Typ I-Fehler. Dieses wird für jedes einzelne Verfahren in den folgenden Abschnitten im Detail erläutert.

### 4.3.1 Haseman-Elston Verfahren

Das HASEMAN-ELSTON Verfahren hält den Typ I-Fehler unter Normalverteilungsannahmen und unabhängiger Geschwisterschaften für alle drei Studiendesigns.

Bei abhängigen Geschwisterschaften zeigt die `Genehunter` Implementation einen nur minimal erhöhten Typ I-Fehler. Einen deutlicheren liberalen Typ I-Fehler zeigt die generalisierte Kleinst-Quadrate-Regression, wie sie in `S.A.G.E.` implementiert ist.

Unter Verletzung der Normalverteilungsannahmen tendiert das HASEMAN-ELSTON Verfahren generell dazu, einen deutlich zu konservativen Typ I-Fehler zu zeigen. Dieser Effekt ist stärker als der gegenläufige Effekt der abhängigen Geschwisterschaften.

Ein Einfluss des Studiendesigns auf den Typ I-Fehler ist nicht zu erkennen.

**Tabelle 11:** Empirischer Typ I-Fehler [in %] des HASEMANN-ELSTON Verfahrens bei einem nominalen Typ I-Fehler von 0,01 sowie zusätzlich unter Einteilung in sieben Kategorien wie in Kapitel 3, Abschnitt 2.2 beschrieben.

| | | Empirischer Typ I-Fehler in % | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | Geschwisterschaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Unter Normalverteilungsannahmen: | | | | | | | | | | |
| Gh.HE.Trad | Unabhängig | 0,966 ○ | 0,992 ○ | 0,966 ○ | 1,008 ○ | 1,018 ○ | 0,963 ○ | 1,014 ○ | 1,010 ○ | 1,080 ○ |
| Sage.HE | | 0,963 ○ | 0,967 ○ | 0,948 ○ | 0,991 ○ | 0,985 ○ | 0,940 ○ | 1,006 ○ | 0,988 ○ | 1,060 ○ |
| Gh.HE.Trad | Abhängig | 1,063 ○ | 1,080 ○ | 1,103 ○ | 1,088 ○ | 1,122 ⇧ | 1,094 ○ | 1,092 ○ | 1,075 ○ | 1,036 ○ |
| Sage.HE | | 1,212 ⇧ | 1,119 ⇧ | 1,286 ⇧ | 1,174 ⇧ | 1,131 ⇧ | 1,167 ⇧ | 1,163 ⇧ | 1,108 ⇧ | 1,089 ○ |
| Unter Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| Gh.HE.Trad | Unabhängig | 0,620 ⇩⇩ | 0,646 ⇩⇩ | 0,620 ⇩⇩ | 0,669 ⇩⇩ | 0,654 ⇩⇩ | 0,683 ⇩⇩ | 0,643 ⇩⇩ | 0,648 ⇩⇩ | 0,622 ⇩⇩ |
| Sage.HE | | 0,672 ⇩⇩ | 0,716 ⇩⇩ | 0,688 ⇩⇩ | 0,735 ⇩⇩ | 0,721 ⇩⇩ | 0,709 ⇩⇩ | 0,728 ⇩⇩ | 0,712 ⇩⇩ | 0,667 ⇩⇩ |
| Gh.HE.Trad | Abhängig | 0,885 ⇩ | 0,865 ⇩ | 0,957 ○ | 0,983 ○ | 0,939 ○ | 0,975 ○ | 0,886 ⇩ | 0,909 ○ | 0,929 ○ |
| Sage.HE | | 1,108 ⇧ | 0,955 ○ | 1,137 ⇧ | 0,732 ⇩⇩ | 0,682 ⇩⇩ | 0,781 ⇩ | 1,169 ⇧ | 0,975 ○ | 1,104 ○ |

### 4.3.2 Revidiertes Haseman-Elston Verfahren

Das REVIDIERTE HASEMAN-ELSTON Verfahren hält den nominalen Typ I-Fehler unter Normalverteilungsannahmen und bei unabhängigen Geschwisterschaften für alle drei Studiendesigns.

Bei abhängigen Geschwisterschaften zeigt sich eine deutliche Inflation des Typ I-Fehlers, die sogar noch stärker ist als bei der generalisierten Kleinst-Quadrate-Regression für das HASEMAN-ELSTON Verfahren.

Unter Verletzung der Normalverteilungsannahmen tendiert das REVIDIERTE HASEMAN-ELSTON Verfahren dazu, schwach konservativ zu werden. Dieser Effekt scheint jedoch wesentlich schwächer ausgeprägt zu sein als beim HASEMAN-ELSTON Verfahren.

Ein Einfluss des Studiendesign auf den Typ I-Fehler ist nicht zu erkennen.

**Tabelle 12:** Empirischer Typ I-Fehler [in %] des REVIDIERTEN HASEMANN-ELSTON Verfahrens bei einem nominalen Typ I-Fehler von 0,01 in % sowie zusätzlich unter Einteilung in sieben Kategorien wie in Kapitel 3, Abschnitt 2.2 beschrieben.

| | | Empirischer Typ I-Fehler in % | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | Geschwisterschaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Unter Normalverteilungsannahmen: | | | | | | | | | | |
| rHE | Unabhängig | 0,990 ○ | 0,998 ○ | 1,009 ○ | 1,024 ○ | 1,044 ○ | 0,932 ○ | 1,021 ○ | 0,989 ○ | 1,018 ○ |
| rHE | Abhängig | 1,662 ⇧⇧ | 1,531 ⇧⇧ | 1,452 ⇧⇧ | 1,356 ⇧⇧ | 1,342 ⇧ | 1386 ⇧⇧ | 1,255 ⇧ | 1,226 ⇧ | 1,210 ⇧ |
| Unter Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| rHE | Unabhängig | 0,971 ○ | 0,879 ⇩ | 0,895 ⇩ | 0,914 ○ | 0,915 ○ | 0,907 ○ | 0,971 ○ | 0,920 ○ | 0,946 ○ |
| rHE | Abhängig | 1,860 ⇧⇧ | 1,700 ⇧⇧ | 1,724 ⇧⇧ | 1,410 ⇧⇧ | 1,319 ⇧ | 1,327 ⇧ | 1,451 ⇧⇧ | 1,328 ⇧ | 1,351 ⇧⇧ |

### 4.3.3 Merlin-Regress Verfahren

MERLIN-REGRESS hält den Typ I-Fehler unter Normalverteilungsannahmen und unabhängiger Geschwisterschaften für alle drei Studiendesigns.

Bei abhängigen Geschwisterschaften zeigt sich eine leichte Inflation des Typ I-Fehlers.

Unter Verletzung der Normalverteilungsannahmen zeigt MERLIN-REGRESS einen konservativen Typ I-Fehler bei unabhängigen Geschwisterschaften, jedoch nicht bei abhängigen Geschwisterschaften. Hier zeigt MERLIN-REGRESS einen liberalen Typ I-Fehler gleicher Größenordnung.

Ein Einfluss des Studiendesign auf den Typ I-Fehler ist nicht zu erkennen.

Da die Anwendung von MERLIN-REGRESS die Schätzung des populationsbasierten Mittelwertes, der Varianz sowie der Heritabilität des Phänotypen voraussetzt, wurde ebenfalls der Effekt einer Ein-Parameter-Misspezifikation überprüft. Die Ergebnisse sind aus Gründen der Übersichtlichkeit im Anhang (Abbildung 6 bis Abbildung 8, S. 73 bis 75) dargestellt. In den Abbildungen ist zu erkennen, dass eine Ein-Parameter-Misspezifikation lediglich zu einem empirischen Powerverlust führt, der Typ I-Fehler jedoch in allen Szenarien nicht beeinflusst wird.

**Tabelle 13:** Empirischer Typ I-Fehler [in %] des MERLIN-REGRESS Verfahrens bei einem nominalen Typ I-Fehler von 0,01 sowie zusätzlich unter Einteilung in sieben Kategorien wie in Kapitel 3, Abschnitt 2.2 beschrieben.

| | | Empirischer Typ I-Fehler in % | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | Geschwisterschaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Unter Normalverteilungsannahmen: | | | | | | | | | | |
| Merlin-Regress | Unabhängig | 0,958 ○ | 0,970 ○ | 0,963 ○ | 1,029 ○ | 1,014 ○ | 0,961 ○ | 0,994 ○ | 1,007 ○ | 0,988 ○ |
| Merlin-Regress | Abhängig | 1,323 ⇧ | 1,256 ⇧ | 1,247 ⇧ | 1,215 ⇧ | 1,248 ⇧ | 1,260 ⇧ | 1,214 ⇧ | 1,195 ⇧ | 1,156 ⇧ |
| Unter Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| Merlin-Regress | Unabhängig | 0,805 ⇩ | 0,811 ⇩ | 0,818 ⇩ | 0,747 ⇩⇩ | 0,816 ⇩ | 0,810 ⇩ | 0,850 ⇩ | 0,848 ⇩ | 0,892 ⇩ |
| Merlin-Regress | Abhängig | 1,293 ⇧ | 1,230 ⇧ | 1,250 ⇧ | 1,185 ⇧ | 1,193 ⇧ | 1,214 ⇧ | 1,212 ⇧ | 1,197 ⇧ | 1,200 ⇧ |

### 4.3.4 Varianzkomponentenmodelle

Die VARIANZKOMPONENTENMODELLE zeigen bei quasi keiner der Szenarien einen korrekten Typ I-Fehler. Sogar für eine zufällig selektierte Stichprobe unter Normalverteilungsannahmen und unabhängiger Geschwisterschaften ist der Typ I-Fehler deutlich zu liberal.

Das Studiendesign der einfachen und doppelten Selektion zeigt einen massiven Einfluss auf den Typ I-Fehler. Während bei der einfachen Selektion die VARIANZKOMPONENTENMODELLE einen deutlich bis massiv zu konservativen Typ I-Fehler zeigen, ist der Typ I-Fehler bei doppelter Selektion massiv zu liberal.

Unter Verletzung der Normalverteilungsannahmen werden die Abweichungen noch drastischer. Abweichungen um den Faktor drei bis fünf des empirischen Typ I-Fehlers vom nominalen Typ I-Fehler treten hier fast immer auf.

Der Faktor der Abweichungen scheint in der Situation der unabhängigen Geschwisterschaften deutlich stärker zu sein als bei den abhängigen Geschwisterschaften.

**Tabelle 14:** Empirischer Typ I-Fehler [in %] der VARIANZKOMPONENTENMODELLE bei einem nominalen Typ I-Fehler von 0,01 sowie zusätzlich unter Einteilung in sieben Kategorien wie in Kapitel 3, Abschnitt 2.2 beschrieben.

| | | Empirischer Typ I-Fehler in % | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | Geschwisterschaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Unter Normalverteilungsannahmen: | | | | | | | | | | |
| Gh.VC | Unabhängig | 1,430 | 1,145 | 1,417 | 0,675 | 0,393 | 0,170 | 2,073 | 2,032 | 2,167 |
| | | ⇧⇧ | ⇧ | ⇧⇧ | ⇩ | ⇩⇩⇩ | ⇩⇩⇩ | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ |
| Solar.VC | | 1,414 | 1,142 | 1,417 | 0,678 | 0,391 | 0,173 | 2,052 | 2,027 | 2,157 |
| | | ⇧⇧ | ⇧ | ⇧⇧ | ⇩ | ⇩⇩⇩ | ⇩⇩⇩ | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ |
| Gh.VC | Abhängig | 1,413 | 1,097 | 1,372 | 1,113 | 0,904 | 0,919 | 1,336 | 1,134 | 1,325 |
| | | ⇧⇧ | ○ | ⇧⇧ | ⇧ | ○ | ○ | ⇧ | ⇧ | ⇧ |
| Solar.VC | | 1,375 | 1,057 | 1,345 | 1,105 | 0,895 | 0,911 | 1,293 | 1,111 | 1,306 |
| | | ⇧⇧ | ○ | ⇧ | ○ | ⇩ | ○ | ⇧ | ⇧ | ⇧ |
| Unter Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| Gh.VC | Unabhängig | 6,260 | 5,749 | 5,163 | 0,365 | 0,256 | 0,100 | 5,486 | 5,685 | 4,880 |
| | | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ | ⇩⇩⇩ | ⇩⇩⇩ | ⇩⇩⇩ | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ |
| Solar.VC | | 6,270 | 5,779 | 5,193 | 0,364 | 0,247 | 0,097 | 5,481 | 5,645 | 4,884 |
| | | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ | ⇩⇩⇩ | ⇩⇩⇩ | ⇩⇩⇩ | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ |
| Gh.VC | Abhängig | 4,764 | 4,130 | 3,979 | 1,525 | 1,432 | 1,060 | 3,947 | 3,748 | 3,427 |
| | | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧ | ⇧⇧ | ○ | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ |
| Solar.VC | | 4,739 | 4,096 | 3,971 | 1,514 | 1,451 | 1,070 | 3,951 | 3,745 | 3,391 |
| | | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧ | ⇧⇧ | ○ | ⇧⇧⇧ | ⇧⇧⇧ | ⇧⇧⇧ |

### 4.3.5 Wilcoxon-Rangsummentest

Der WILCOXON-RANGSUMMENTEST hält den korrekten Typ I-Fehler unter Normalverteilungsannahmen sowie unter Verletzung der Normalverteilungsannahmen bei unabhängiger Geschwisterschaften für alle drei Studiendesigns.

Bei abhängigen Geschwisterschaften zeigt sich eine leichte aber deutliche Inflation des Typ I-Fehlers.

Ein Einfluss des Studiendesigns auf den Typ I-Fehler ist nicht zu erkennen.

**Tabelle 15:** Empirischer Typ I-Fehler [in %] des WILCOXON-RANGSUMMENTESTS bei einem nominalen Typ I-Fehler von 0,01 sowie zusätzlich unter Einteilung in sieben Kategorien wie in Kapitel 3, Abschnitt 2.2 beschrieben.

| | | Empirischer Typ I-Fehler in % | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | Geschwisterschaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Unter Normalverteilungsannahmen: | | | | | | | | | | |
| Npar | Unabhängig | 0,989 | 0,961 | 0,929 | 0,972 | 0,954 | 1,075 | 0,978 | 0,969 | 0,967 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Npar | Abhängig | 1,142 | 1,195 | 1,144 | 1,065 | 1,196 | 1,156 | 1,131 | 1,135 | 1,134 |
| | | ⇧ | ⇧ | ⇧ | ○ | ⇧ | ⇧ | ⇧ | ⇧ | ⇧ |
| Unter Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| Npar | Unabhängig | 0,989 | 0,961 | 0,929 | 0,972 | 0,954 | 1,075 | 0,978 | 0,969 | 0,967 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Npar | Abhängig | 1,106 | 1,179 | 1,116 | 1,208 | 1,142 | 1,218 | 1,132 | 1,080 | 1,118 |
| | | ⇧ | ⇧ | ⇧ | ⇧ | ⇧ | ⇧ | ⇧ | ○ | ⇧ |

### 4.3.6 Merlin-QTL Verfahren

Die Kong und Cox Teststatistik (1997) des MERLIN-QTL Verfahrens hält unter allen Simulationen das korrekte Typ I-Fehlerniveau. Es ist kein Einfluss von Selektion, Verteilungsannahmen oder abhängiger Geschwisterschaften auf den Typ I-Fehler zu erkennen.

Die Whittemore und Halpern Teststatistik (1994) des MERLIN-QTL Verfahrens zeigt unter allen Simulationen einen deutlich zu konservativen Typ I-Fehler. Es ist kein Einfluss von Selektion, Verteilungsannahmen oder abhängiger Geschwisterschaften auf den Typ I-Fehler zu erkennen.

**Tabelle 16:** Empirischer Typ I-Fehler [in %] des MERLIN-QTL Verfahrens bei einem nominalen Typ I-Fehler von 0,01 sowie zusätzlich unter Einteilung in sieben Kategorien wie in Kapitel 3, Abschnitt 2.2 beschrieben.

| | | Empirischer Typ I-Fehler in % | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | Geschwisterschaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Unter Normalverteilungsannahmen: | | | | | | | | | | |
| Merlin.K&C | Unabhängig | 0,994 | 0,996 | 0,966 | 0,953 | 0,929 | 0,919 | 0,954 | 0,994 | 0,960 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Merlin.W&H | | 0,674 | 0,690 | 0,688 | 0,644 | 0,643 | 0,610 | 0,649 | 0,690 | 0,664 |
| | | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ |
| Merlin.K&C | Abhängig | 0,954 | 1,022 | 0,938 | 0,930 | 0,995 | 0,979 | 0,930 | 0,999 | 0,908 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Merlin.W&H | | 0,721 | 0,765 | 0,681 | 0,669 | 0,729 | 0,734 | 0,699 | 0,725 | 0,667 |
| | | ⇩⇩ | ⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ |
| Unter Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| Merlin.K&C | Unabhängig | 0,957 | 1,019 | 0,945 | 0,961 | 0,970 | 0,966 | 1,009 | 0,950 | 1,007 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Merlin.W&H | | 0,666 | 0,704 | 0,634 | 0,665 | 0,649 | 0,673 | 0,711 | 0,629 | 0,687 |
| | | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ |
| Merlin.K&C | Abhängig | 0,941 | 0,971 | 0,946 | 0,938 | 0,927 | 0,945 | 0,946 | 0,997 | 1,011 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Merlin.W&H | | 0,734 | 0,712 | 0,724 | 0,695 | 0,673 | 0,697 | 0,720 | 0,744 | 0,750 |
| | | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ | ⇩⇩ |

### 4.3.7 Maximum Likelihood Binomial Verfahren

Das MAXIMUM LIKELIHOOD BINOMIAL Verfahren hält sowohl unter der Verwendung der Normalverteilungsannahmen, als auch unter der Verwendung der empirischen Verteilungsfunktion mit zehn Kategorien unter allen Simulationen das korrekte Typ I-Fehlerniveau. Es ist kein Einfluss von Selektion, Verteilungsannahmen oder abhängiger Geschwisterschaften auf den Typ I-Fehler zu erkennen.

**Tabelle 17:** Empirischer Typ I-Fehler [in %] des MAXIMUM LIKELIHOOD BINOMIAL Verfahrens bei einem nominalen Typ I-Fehler von 0,01 sowie zusätzlich unter Einteilung in sieben Kategorien wie in Kapitel 3, Abschnitt 2.2 beschrieben.

| | | Empirischer Typ I-Fehler in % | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | Geschwisterschaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Unter Normalverteilungsannahmen: | | | | | | | | | | |
| Mlbqt.NV | Unabhängig | 1,065 | 1,134 | 1,056 | 1,058 | 1,063 | 0,956 | 1,018 | 1,086 | 1,033 |
| | | ○ | ⇧ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Mlbqt.Kat | | 1,068 | 1,131 | 1,044 | 1,015 | 1,081 | 0,981 | 1,022 | 1,084 | 1,013 |
| | | ○ | ⇧ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Mlbqt.NV | Abhängig | 0,989 | 0,986 | 0,970 | 0,987 | 1,011 | 1,037 | 0,932 | 0,996 | 0,949 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Mlbqt.Kat | | 1,000 | 1,030 | 0,987 | 0,971 | 0,981 | 1,057 | 0,938 | 1,008 | 0,931 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Unter Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| Mlbqt.NV | Unabhängig | 1,044 | 1,053 | 1,082 | 1,077 | 1,008 | 0,986 | 1,037 | 0,967 | 1,055 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Mlbqt.Kat | | 1,043 | 1,025 | 1,038 | 1,043 | 0,975 | 1,009 | 1,039 | 0,964 | 1,014 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Mlbqt.NV | Abhängig | 0,983 | 0,966 | 1,049 | 0,985 | 0,996 | 0,985 | 1,024 | 0,978 | 1,007 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Mlbqt.Kat | | 0,983 | 0,966 | 1,049 | 0,985 | 0,996 | 0,985 | 1,024 | 0,978 | 1,007 |
| | | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

### 4.3.8 Modellbasiertes Verfahren

Das voll parametrisierte MODELLBASIERTE Verfahren zeigt einen deutlich bis massiv konservativen Typ I-Fehler unter fast allen Simulationen sogar unter Normalverteilungsannahmen und unabhängiger Geschwisterschaften für alle drei Studiendesigns.

Der Effekt von abhängigen Geschwisterschaften und Verletzung der Normalverteilungsannahmen auf den Typ I-Fehler ist unklar.

Unter einfacher und doppelter Selektion scheint der Typ I-Fehler im Vergleich zu den Simulationen unter zufälliger Selektion anzusteigen.

**Tabelle 18:** Empirischer Typ I-Fehler [in %] des MODELLBASIERTEN Verfahrens bei einem nominalen Typ I-Fehler von 0,01 sowie zusätzlich unter Einteilung in sieben Kategorien wie in Kapitel 3, Abschnitt 2.2 beschrieben.

| | | Empirischer Typ I-Fehler in % | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | Geschwisterschaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Unter Normalverteilungsannahmen: | | | | | | | | | | |
| Linkage | Unabhängig | 0,399 | 0,014 | 0,170 | 1,040 | 0,229 | 0,823 | 1,049 | 0,485 | 0,920 |
| | | ⇩⇩⇩ | ⇩⇩⇩ | ⇩⇩⇩ | ○ | ⇩⇩⇩ | ⇩ | ○ | ⇩⇩⇩ | ○ |
| Linkage | Abhängig | 0,483 | 0,015 | 0,236 | 0,915 | 0,147 | 0,702 | 0,812 | 0,139 | 0,593 |
| | | ⇩⇩⇩ | ⇩⇩⇩ | ⇩⇩⇩ | ○ | ⇩⇩⇩ | ⇩⇩ | ⇩ | ⇩⇩⇩ | ⇩⇩ |
| Unter Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| Linkage | Unabhängig | 0,580 | 0,021 | 0,288 | 1,062 | 0,304 | 1,059 | 1,081 | 0,368 | 0,932 |
| | | ⇩⇩ | ⇩⇩⇩ | ⇩⇩⇩ | ○ | ⇩⇩⇩ | ○ | ○ | ⇩⇩⇩ | ○ |
| Linkage | Abhängig | 0,593 | 0,031 | 0,360 | 0,984 | 0,213 | 0,861 | 0,882 | 0,159 | 0,659 |
| | | ⇩⇩ | ⇩⇩⇩ | ⇩⇩⇩ | ○ | ⇩⇩⇩ | ⇩ | ⇩ | ⇩⇩⇩ | ⇩⇩ |

## 4.4 Empirischer Powervergleich der Verfahren

Die Ergebnisse der Monte-Carlo Simulationen zur Ermittlung der empirischen Power sind in der Tabelle 19 für die Simulationen unter Normalverteilungsannahmen sowie in Tabelle 20 für Simulationen unter der Verletzung der Normalverteilungsannahmen gezeigt.

Im Allgemeinen ist zu erkennen, dass das voll parametrisierte MODELLBASIERTE Verfahren für die dominanten und rezessiven Modelle die höchste Power zeigt. Aufgrund der verwendeten wahren, aber in realen Studien quasi unschätzbaren Modellspezifikation für den QTL war dieses Ergebnis zu erwarten (siehe Kapitel 2, Abschnitt 8, S. 17). Das MODELLBASIERTE Verfahren kann deshalb als eine Art Goldstandard aufgefasst werden, das jedoch bei realen Studien in der Regel nicht verwendet werden kann. Das MODELLBASIERTE Verfahren wird deshalb bei den weiteren Ausführungen zum Powervergleich nicht weiter betrachtet.

**Tabelle 19**: Empirische Power [%] bei einem empirischen Fehlerniveau von 5% unter Normalverteilungsannahmen.

| | | Empirische Power unter Normalverteilungsannahmen | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Gh.HE.Trad | Unabhängig | 34,3 | 38,4 | 32,4 | 69,2 | 64,3 | 64,0 | 58,7 | 66,7 | 57,2 |
| Sage.HE | | 34,0 | 38,0 | 32,3 | 68,9 | 63,9 | 63,9 | 58,7 | 66,8 | 57,3 |
| rHE | | 24,0 | 27,6 | 26,4 | 59,6 | 55,9 | 60,2 | 62,7 | 59,7 | 62,4 |
| Merlin-Regress | | 41,9 | 48,9 | 39,7 | 79,0 | 76,2 | 79,3 | 75,8 | 76,7 | 79,6 |
| Gh.VC | | 43,5 | 47,4 | 41,7 | 82,0 | 71,7 | 77,3 | 73,0 | 76,2 | 79,0 |
| Solar.VC | | 43,5 | 47,4 | 41,5 | 82,0 | 71,8 | 77,6 | 72,9 | 76,7 | 78,9 |
| Npar | | 20,9 | 27,2 | 22,7 | 60,1 | 57,9 | 57,9 | 36,2 | 44,1 | 37,6 |
| Merlin.K&C | | 13,1 | 18,7 | 13,5 | 40,2 | 40,2 | 43,0 | 45,7 | 53,9 | 45,2 |
| Merlin.W&H | | 15,1 | 20,4 | 14,5 | 43,1 | 42,2 | 44,1 | 46,8 | 55,8 | 47,1 |
| Mlbqt.N | | 21,6 | 23,6 | 20,9 | 46,1 | 46,0 | 44,3 | 51,3 | 54,4 | 51,2 |
| Mlbqt.Kat | | 21,0 | 22,7 | 18,7 | 43,5 | 45,3 | 41,2 | 49,1 | 54,1 | 48,5 |
| Linkage | | 69,0 | 40,5 | 57,8 | 94,4 | 65,5 | 89,8 | 91,7 | 70,4 | 88,6 |
| Gh.HE.Trad | Abhängig | 39,6 | 43,4 | 35,8 | 65,2 | 59,7 | 57,1 | 51,4 | 56,3 | 47,8 |
| Sage.HE | | 43,6 | 45,3 | 40,7 | 70,0 | 64,4 | 62,3 | 54,8 | 59,3 | 54,1 |
| rHE | | 35,8 | 41,5 | 39,2 | 64,0 | 57,4 | 58,5 | 52,0 | 56,7 | 52,3 |
| Merlin-Regress | | 49,3 | 52,3 | 47,1 | 77,4 | 71,9 | 72,1 | 62,8 | 66,2 | 64,2 |
| Gh.VC | | 52,0 | 53,9 | 48,8 | 77,8 | 70,2 | 72,6 | 63,0 | 65,6 | 64,9 |
| Solar.VC | | 51,3 | 53,4 | 48,4 | 77,3 | 70,8 | 73,0 | 62,9 | 66,7 | 65,0 |
| Npar | | 23,7 | 30,9 | 26,4 | 55,7 | 48,8 | 49,5 | 38,7 | 43,4 | 36,8 |
| Merlin.K&C | | 15,9 | 23,0 | 18,2 | 43,0 | 40,0 | 35,8 | 30,0 | 40,7 | 28,0 |
| Merlin.W&H | | 18,6 | 25,4 | 19,6 | 46,9 | 42,3 | 39,0 | 32,2 | 43,9 | 30,0 |
| Mlbqt.N | | 22,5 | 28,6 | 22,5 | 45,2 | 43,9 | 35,7 | 34,8 | 41,7 | 32,6 |
| Mlbqt.Kat | | 23,1 | 28,1 | 21,3 | 43,6 | 42,6 | 32,9 | 33,3 | 40,7 | 31,1 |
| Linkage | | 69,3 | 48,9 | 59,4 | 88,8 | 64,2 | 80,2 | 83,2 | 62,2 | 75,2 |

**Tabelle 20**: Empirische Power [%] bei einem empirischen Fehlerniveau von 5% unter Verletzung der Normalverteilungsannahmen.

| | | Empirische Power unter Verletzung der Normalverteilungsannahmen | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Gh.HE.Trad | Unabhängig | 18,3 | 17,9 | 17,4 | 16,1 | 10,9 | 15,7 | 17,4 | 20,9 | 19,1 |
| Sage.HE | | 18,1 | 17,8 | 17,1 | 15,7 | 11,0 | 15,7 | 17,4 | 21,0 | 19,2 |
| rHE | | 28,0 | 31,5 | 28,7 | 59,5 | 47,3 | 55,1 | 41,7 | 29,4 | 46,8 |
| Merlin-Regress | | 41,1 | 38,3 | 35,1 | 53,7 | 40,5 | 53,8 | 50,2 | 51,0 | 58,4 |
| Gh.VC | | 40,1 | 33,8 | 34,6 | 44,5 | 25,4 | 42,6 | 40,8 | 42,5 | 55,2 |
| Solar.VC | | 40,4 | 34,6 | 34,7 | 44,2 | 25,5 | 43,2 | 40,7 | 42,3 | 55,2 |
| Npar | | 40,4 | 83,4 | 34,5 | 82,7 | 70,7 | 77,2 | 36,9 | 62,9 | 35,5 |
| Merlin.K&C | | 25,1 | 47,9 | 23,6 | 76,1 | 55,6 | 72,8 | 55,6 | 72,8 | 51,9 |
| Merlin.W&H | | 25,8 | 50,0 | 25,1 | 77,8 | 59,0 | 73,1 | 57,7 | 75,3 | 54,1 |
| Mlbqt.N | | 29,7 | 32,9 | 27,2 | 68,4 | 52,6 | 60,5 | 55,6 | 67,4 | 56,7 |
| Mlbqt.Kat | | 29,6 | 36,1 | 24,0 | 64,3 | 51,9 | 51,0 | 54,7 | 69,3 | 55,0 |
| Linkage | | 65,6 | 38,7 | 44,6 | 95,3 | 54,2 | 73,2 | 76,4 | 59,6 | 69,0 |
| Gh.HE.Trad | Abhängig | 21,7 | 26,5 | 23,4 | 26,7 | 22,3 | 24,8 | 29,6 | 23,2 | 27,5 |
| Sage.HE | | 36,3 | 39,9 | 36,0 | 44,4 | 36,8 | 44,5 | 45,8 | 39,2 | 43,8 |
| rHE | | 43,6 | 51,1 | 45,8 | 65,3 | 59,8 | 64,9 | 61,9 | 58,9 | 59,1 |
| Merlin-Regress | | 49,5 | 54,1 | 48,9 | 64,2 | 52,4 | 63,2 | 62,9 | 58,3 | 62,4 |
| Gh.VC | | 43,4 | 48,8 | 46,1 | 57,3 | 45,1 | 56,2 | 57,4 | 48,1 | 59,2 |
| Solar.VC | | 43,4 | 49,8 | 46,4 | 57,6 | 45,0 | 56,6 | 57,9 | 48,6 | 58,9 |
| Npar | | 42,5 | 86,2 | 40,7 | 81,0 | 86,9 | 74,7 | 61,4 | 88,3 | 57,8 |
| Merlin.K&C | | 24,7 | 51,8 | 24,5 | 65,5 | 71,9 | 58,8 | 48,1 | 64,2 | 41,8 |
| Merlin.W&H | | 27,3 | 55,7 | 27,0 | 68,6 | 74,5 | 62,7 | 51,7 | 67,1 | 46,8 |
| Mlbqt.N | | 30,1 | 40,0 | 31,9 | 60,8 | 56,3 | 56,0 | 49,8 | 52,7 | 46,8 |
| Mlbqt.Kat | | 30,0 | 43,2 | 29,0 | 59,3 | 57,0 | 48,7 | 48,0 | 54,2 | 42,9 |
| Linkage | | 66,3 | 51,1 | 51,1 | 89,1 | 61,8 | 71,5 | 81,3 | 57,2 | 67,0 |

### 4.4.1 Empirischer Powervergleich innerhalb der Verfahren

Alle Verfahren zeigen unter Normalverteilungsannahmen bei einfacher oder doppelter Selektion eine höhere empirische Power im Vergleich zur populationsbasierten Stichprobe. Hingegen zeigen bei Verletzung der Normalverteilungsannahmen lediglich das REVIDIERTE HASEMAN-ELSTON Verfahren, MERLIN-REGRESS, der WILCOXON-RANGSUMMENTEST sowie die Allele-sharing Ansätze MERLIN-QTL und das MAXIMUM LIKELIHOOD BINOMIAL Verfahren eine höhere empirische Power im Vergleich zum populationsbasierten Studiendesign. Beim HASEMAN-ELSTON Verfahren und den VARIANZKOMPONENTENMODELLEN ist, wenn überhaupt, nur sporadisch eine höhere empirische Power zu erkennen.

Der direkte Powervergleich der einzelnen Verfahren unter Normalverteilungsannahmen gegenüber den ansonsten gleichen Modellen unter Verletzung der Normalverteilungsannahmen ergibt ein heterogenes Bild, welches nachfolgend erläutert wird.

Beim HASEMAN-ELSTON Verfahren ist unter Verletzung der Normalverteilungsannahmen ein deutlicher Einbruch der empirischen Power zu erkennen, der in den nicht populationsbasierten

Studiendesigns sogar dazu führt, dass gegenüber der zufälligen Selektion überhaupt kein Power-Gewinn mehr erkennbar ist. Während sich unter zufälliger Selektion bei MERLIN-REGRESS und den VARIANZKOMPONENTENMODELLEN nur eine leicht geringere empirischen Power zeigt, ist gerade bei den VARIANZKOMPONENTENMODELLEN ein deutlicher Einbruch der Power in den nicht populationsbasierten Studiendesigns zu erkennen. Das REVIDIERTE HASEMAN-ELSTON Verfahren zeigt unter Verletzung der Normalverteilungsannahmen eine sehr robuste Power, d.h. ein Einfluss auf die empirische Power ist nicht erkennbar.
Der WILCOXON-RANGSUMMENTEST sowie die Allele-sharing Verfahren - MERLIN-QTL und das MAXIMUM LIKELIHOOD BINOMIAL Verfahren - zeigen in allen drei Studiendesigns eine deutlich höhere empirische Power.

### 4.4.2 Empirische Power der Verfahren im direkten Vergleich

Zunächst werden die Ergebnisse des Vergleichs der Verfahren unter Normalverteilungsannahmen erläutert. MERLIN-REGRESS sowie die VARIANZKOMPONENTENMODELLE zeigen generell eine hohe empirische Power. Die beiden anderen regressionsbasierten Verfahren, nämlich das HASEMAN-ELSTON und das REVIDIERTE HASEMAN-ELSTON Verfahren folgen mit einem deutlichen Abstand. Die niedrigste Power erreichen die Allele-sharing Verfahren MERLIN-QTL und das MAXIMUM LIKELIHOOD BINOMIAL Verfahren sowie der WILCOXON-RANGSUMMENTEST. Diese Reihenfolge ist unter allen drei Studiendesigns gleich.
Die generalisierte Kleinst-Quadrate-Regression beim HASEMAN-ELSTON Verfahren zeigt bei abhängigen Geschwisterschaften eine leicht höhere empirische Power, jedoch ist die empirische Power immer noch deutlich niedriger als bei MERLIN-REGRESS sowie den VARIANZKOMPONENTENMODELLEN.
Unter Verletzung der Normalverteilungsannahmen ergeben sich für die Studiendesigns z. T. unterschiedliche Resultate. Deshalb werden die Ergebnisse für jedes Verfahren einzeln erläutert.
Das HASEMAN-ELSTON Verfahren zeigt in allen drei Studiendesigns mit Abstand die niedrigste empirische Power. Die generalisierte Kleinst-Quadrate-Regression des HASEMAN-ELSTON Verfahren verbessert die empirische Power bei abhängigen Geschwisterschaften - speziell in dem populationsbasierten Studiendesign -, jedoch reicht die Power nur in Einzelfällen an die Power der anderen Verfahren heran.
Das REVIDIERTE HASEMAN-ELSTON Verfahren zeigt bei allen Modellen eine stabile mittlere empirische Power. Bei abhängigen Geschwisterschaften schneidet es deutlich besser ab.
Die VARIANZKOMPONENTENMODELLE und MERLIN-REGRESS zeigen unter dem populationsbasierten Studiendesign eine hohe Power. Unter einfacher und doppelter Selektion zeigen jedoch MERLIN-QTL und insbesondere der WILCOXON-RANGSUMMENTEST die höchste Power. Die

VARIANZKOMPONENTENMODELLE haben unter allen Modellen eine niedrigere empirische Power als MERLIN-REGRESS.

Der WILCOXON-RANGSUMMENTEST zeigt fast immer die höchste Power. Eine systematische Ausnahme ist lediglich unter doppelter Selektion bei unabhängigen Geschwisterschaften zu erkennen.

Für MERLIN-QTL und das MAXIMUM LIKELIHOOD BINOMIAL Verfahren ergab sich eine mittlere Power in einer populationsbasierten Stichprobe, jedoch sind diese beiden Verfahren in den Studiendesign unter Selektion z.T. vergleichbar mit der hohen Power des WILCOXON-RANGSUMMENTEST.

### 4.4.3 Zusammenfassung der empirischen Powervergleiche

HASEMAN-ELSTON Verfahren

Das HASEMAN-ELSTON Verfahren hat unter Normalverteilungsannahmen unabhängig vom Studiendesign eine mittlere Power. Unter Verletzung der Normalverteilungsannahmen führt der drastische Einbruch der emprirische Power generell dazu, dass es die geringste Power zeigt.

Bei abhängigen Geschwisterschaften ist die empirische Power bei der generalisierten Kleinst-Quadrate-Regression gegenüber der gewöhnlichen Kleinst-Quadrate-Regression erhöht, dies hat jedoch auf die Rangfolge keinen Einfluss hat.

Für alle betrachteten Szenarien ist das HASEMAN-ELSTON Verfahren MERLIN-REGRESS unterlegen.

REVIDIERTE HASEMAN-ELSTON Verfahren

Das REVIDIERTE HASEMAN-ELSTON Verfahren zeigt unabhängig von Studiendesign oder Verteilungsannahmen durchweg eine robuste mittlere Power.

MERLIN-REGRESS

Unter Normalverteilungsannahmen zeigt MERLIN-REGRESS unabhängig vom Studiendesign eine hohe mit den VARIANZKOMPONENTENMODELLEN vergleichbare empirische Power. Unter Verletzung der Normalverteilungsannahmen ist es den VARIANZKOMPONENTENMODELLEN jedoch überlegen. Hier ist die Power ebenfalls hoch, jedoch sind MERLIN-QTL und insbesondere der WILCOXON-RANGSUMMENTEST besonders unter Selektion überlegen.

Eine Ein-Parameter-Misspezifikation führt lediglich zu einem empirischen Powerverlust, jedoch wird der Typ I-Fehler in allen Szenarien nicht beeinflusst (Abbildung 6 bis Abbildung 8, S. 73 bis 75). Eine Misspezifikation des Mittelwertes zeigte den größten Einfluss auf die Power. Ein Überschätzen der Varianz hat einen schwächeren jedoch immer noch starken Effekt, während ein Unterschätzen der Varianz keinen Powerverlust nach sich zu ziehen scheint. Der Effekt einer

moderaten und realistischen Misspezifikation der Heritabilität ist gering; er wird jedoch größer, wenn man sich der oberen und unteren Grenze (0 und 1) nähert.

VARIANZKOMPONENTENMODELLE

Unter Normalverteilungsannahmen zeigen die VARIANZKOMPONENTENMODELLE unabhängig vom Studiendesign zusammen mit MERLIN-REGRESS die höchste Power. Unter Verletzung der Normalverteilungsannahmen ist die Power für das populationsbasierte Studiendesign ebenfalls hoch, jedoch ist ein deutlicher Einbruch der Power unter Selektion zu erkennen. Der WILCOXON-RANGSUMMENTEST sowie die Allele-sharing Verfahren haben dann im Allgemeinen eine höhere Power.

Die Ergebnisse der beiden betrachteten Software-Implementationen in `Solar` und `Genehunter` legen keine Unterschiede in der empirischen Power nahe.

WILCOXON-RANGSUMMENTEST

Der WILCOXON-RANGSUMMENTEST zeigt unter Normalverteilungsannahmen unabhängig vom Studiendesign eine geringe empirische Power. Unter Verletzung der Normalverteilungsannahmen ändert sich jedoch die Situation. Während unter dem populationsbasierten Studiendesign die Power vergleichbar mit denen von MERLIN-REGRESS sowie den VARIANZKOMPONENTENMODELLE ist, zeigt der WILCOXON-RANGSUMMENTEST unter den beiden selektierten Studiendesigns fast immer die höchste Power. Eine systematische Ausnahme ist lediglich unter doppelter Selektion bei unabhängigen Geschwisterschaften zu erkennen. Darüber hinaus ist erwähnenswert, dass unter Verletzung der Normalverteilungsannahmen und zufälliger Selektion die Power im additiven Model im Vergleich zu dem dominanten und rezessiven Model deutlich höher ist.

MERLIN-QTL und MAXIMUM LIKELIHOOD BINOMIAL Verfahren

Beide Verfahren zeigen unter Normalverteilungsannahmen unabhängig vom Studiendesign eine niedrige empirische Power. Allerdings sind diese beiden Verfahren in den Studiendesigns unter Selektion z.T. vergleichbar mit der hohen Power des WILCOXON-RANGSUMMENTESTS.

## 4.5 Analyse des COAG Perth Datensatzes

Die im Rahmen dieser Arbeit verwendeten kopplungsanalytischen Verfahren sollen nun durch die Anwendung auf einen realen Datensatz illustriert werden. Hierbei wird insbesondere Wert darauf gelegt, die praktische Vorgehensweise der genetischen Kartierung eines quantitativen Merkmals näher zu erläutern.

Hierzu wurde der Datensatz „Consortium on Asthma Genetics: Perth study“ (COAG Perth Datensatz), wie er für den Genetic Analysis Workshop 12 zur Verfügung gestellt wurde, erneut analysiert.

Im ersten Schritt wurden zunächst die Daten des Original-Datensatzes so aufbereitet, dass sie von der von Herrn Dr. Franke und dem Verfasser erstellten Software `Abi2Link` verwendet werden konnten. Die Formatierung der Eingabedateien lehnt sich dabei an das Ausgabeformat der meisten Genotypisierungsplattformen an. Deshalb waren nur minimale Anpassungen wie z.B. die Änderung der Dateinamen notwendig. Die Eingangsformate sind inkl. Beispieldateien im Softwarepaket dokumentiert und deshalb hier nicht näher erläutert. `Abi2Link` diente dann zur automatischen skriptgesteuerten Erstellung der notwendigen Dateien im `Linkage`-Datenformat, welche im Folgenden verwendet wurden. Das Programm `Abi2Link` überprüft dabei automatisch eine Reihe von logischen Fehlern in den Daten und erstellt ein Protokoll. Fehler dieser Art können z.B. sein, dass Genotypen an einem Marker für eine Person mehrfach im Datensatz vorhanden sind oder Familien mehr als einmal in das `Linkage`-Ausgabeformat geschrieben werden sollen. `Abi2Link` dient also gleichzeitig einer ersten Überprüfung der Daten auf logische Fehler. Da die Eltern genotypisiert wurden, konnten die Allelfrequenzen der Marker aus den Eltern mit Hilfe von `Abi2Link` gleichzeitig geschätzt und später für die Analysen verwendet werden.

Dann wurden die im Datensatz erhaltenen Genotypen mit dem Programm `Pedcheck` auf Vererbungsfehler überprüft (O'Connell und Weeks, 1998). Genotypen, welche die Vererbungsregeln verletzen wurden von den weiteren Untersuchungen ausgeschlossen. Es gab Aufgrund der Vererbungsregeln verletzenden Genotypen keine Hinweise auf Fehler in den Familienstrukturen, wie z.B. fehlende zusätzliche Väter. Ebenfalls wurden Familien ohne Phäno- oder Genotypen unter den Kindern von der Analyse ausgeschlossen, da sie keinen Beitrag zur Kopplungsanalyse haben. `Pedcheck` ergab ebenfalls keine Hinweise auf Verletzung des Hardy-Weinberg-Gleichgewichtes der beiden Marker (Ergebnisse hier nicht im Detail gezeigt).

Nach der Qualitätskontrolle standen 82 Familien von ursprünglich 98 Familien mit 195 Kindern zur Verfügung, die sowohl Phänotypen als auch Genotypen enthielten. Davon bestanden 56 Familien aus Eltern und zwei Kindern, während die restlichen 26 Familien zwischen drei und vier Kinder pro Familie aufwiesen.

Die Abbildung 5 zeigt die Verteilung der Phänotypen lnIge und lnIgeR der Kinder in einem Histogramm. Während beide Histogramme keine Ausreißer zeigen, scheint jedoch im visuellen Vergleich der Phänotyp ln IgER besser einer Normalverteilung zu entsprechen. Der Phänotyp wird hier in seiner logarithmierten Form verwendet, da diese besser zu einer Normalverteilung zu passen scheint. Diese Transformation der Daten ist ein durchaus übliches Vorgehen.

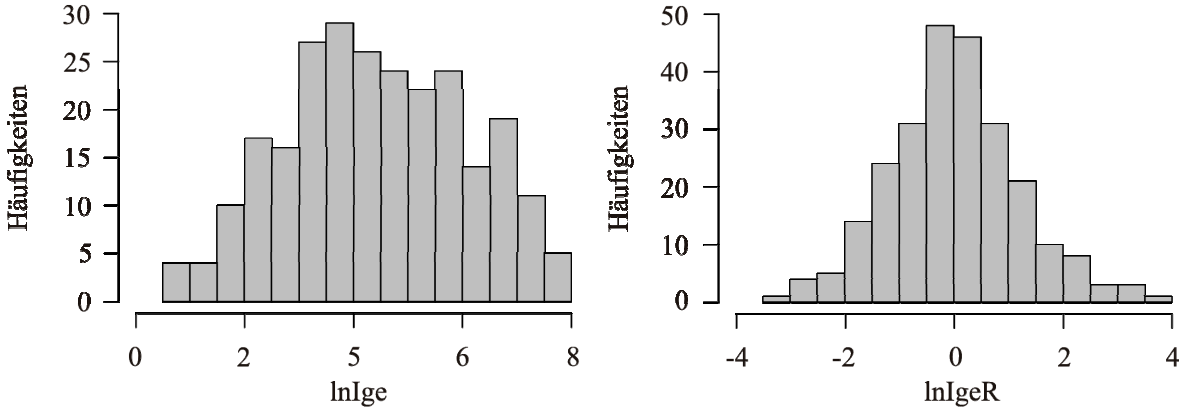


**Abbildung 5**: Histogramme der Kinder des COAG Perth Datensatzes für den logarithmierten Gesamt Serum IgE Titer (ln IgE) sowie adjustierten logarithmierte Gesamt Serum IgE Titer (ln IgER).

Nach der Qualitätskontrolle und Aufbereitung der Daten wurden nun die acht Verfahren mit den in Kapitel 3 Abschnitt 2.1 beschriebenen Analyseoptionen für eine Multipoint-Kopplungsanalyse angewandt. Mittelwert, Varianz und empirische Dezile wurden hierbei, soweit erforderlich, aus den Phänotypen der Kinder geschätzt. Für Merlin-Regress wurde zusätzlich der Standardparameter für die Heritabilität von 0,5 verwendet. Soweit im Softwarepaket möglich, wurden zusätzlich empirische *p*-Werte durch 100.000 Permutationen ermittelt.

**Tabelle 21**: Asymptotische und empirische *p*-Werte der Anwendung der kopplungsanalytischen Verfahren auf den COAG Perth Datensatz.

| | ln IgE | | ln IgER | |
|---|---|---|---|---|
| Verfahren | D5S393 | D5S399 | D5S393 | D5S399 |
| Asymptotische *p* -Werte: | | | | |
| Gh.HE.trad | 0,0660 | 0,0751 | 0,0441 | 0,0126 |
| Sage.HE | 0,0648 | 0,0739 | 0,0310 | 0,0096 |
| rHe | 0,2831 | 0,3002 | 0,0047 | 0,0013 |
| Merlin-Regress | 0,0110 | 0,0300 | 0,0004 | 0,0004 |
| Gh.VC | 0,0151 | 0,0223 | 0,0032 | 0,0023 |
| Solar.VC | 0,0083 | 0,0084 | 0,0032 | 0,0023 |
| Npar | 0,0528 | 0,0544 | 0,0805 | 0,0540 |
| Merlin.K&C | 0,3632 | 0,3336 | 0,1239 | 0,0173 |
| Merlin.W&H | 0,1251 | 0,1003 | 0,1515 | 0,0314 |
| Mlbqt.Kat | 0,0379 | 0,0391 | 0,0062 | 0,0048 |
| Empirische *p* -Werte: | | | | |
| Sage.HE | 0,0492 | 0,0599 | 0,0411 | 0,0121 |
| rHE | 0,3204 | 0,2982 | 0,0041 | 0,0008 |
| Merlin-Regress | 0,0201 | 0,0383 | 0,0018 | 0,0016 |

Auf die Anwendung der MODELLBASIERTEN Kopplungsanalyse wurde verzichtet, da die dafür notwendige Schätzung der Parameter aus dem Datensatz nicht möglich ist. Ebenso wurde auf das MAXIMUM LIKELIHOOD BINOMIAL Verfahren unter Normalverteilungsannahmen verzichtet, da

zumindest der der Phänotyp ln IgER nicht einer Standardnormalverteilung folgt. In Tabelle 21 sind die Ergebnisse der Kopplungsanalysen als Übersicht zusammengefasst.

Generell ist zu erkennen, dass die *p*-Werte für den gegen Alter, Geschlecht und Antigen spezifischen IgE Titer korrigierten logarithmierten IgE Titer kleiner sind als für den unkorrigierten logarithmierten IgE Titer. Bemerkenswert sind die teilweise massiven Unterschiede der *p*-Werte im Methodenvergleich. Während für den korrigierten logarithmierten IgE Titer der höchste asymptotische *p*-Wert am Marker D5S393 unter Verwendung des MERLIN-QTL Verfahrens mit der Whittemore und Halpern Teststatistik bei 15% liegt, ist der niedrigste asymptotische *p*-Wert 0,04% bzw. 0,16% als empirischer *p*-Wert unter Verwendung von MERLIN-REGRESS.

# 5 Diskussion

Ziel dieser Arbeit war ein umfassender Gütevergleich häufig verwendeter kopplungsanalytischer Verfahren zur genetischen Kartierung quantitativer Merkmale. Zum ersten Mal wurden die Güteeigenschaften dieser Verfahren in einem solchem Umfang unter einer Vielzahl realistischer Annahmen und Studiendesigns in einem direkten Vergleich in einer Monte-Carlo Simulationsstudie gegenübergestellt.

Nachfolgend wird zunächst im ersten Abschnitt das gewählte Simulationsmodell diskutiert. Im nachfolgenden Abschnitt wird dann die Simulationssoftware `Sibsim` sowie mögliche Anwendung auf weitere Fragestellungen näher betrachtet. Der dritte Abschnitt untersucht und diskutiert die Güteeigenschaften der kopplungsanalytischen Verfahren, wie sie im Rahmen dieser Arbeit ermittelt wurden. In den beiden letzten Abschnitten werden dann die Ergebnisse der erneuten Analyse des COAG Perth Datensatzes diskutiert sowie abschließend ein Ausblick gegeben.

## 5.1 Ausgewählte Simulationsmodelle und Verfahren

Zur Durchführung dieses Vergleiches wurde zunächst ein für die Simulationen des QTL und der Phänotypen geeignetes Modell gewählt. Das additive Model von Falconer und Mackay (1996) schien auch schon aufgrund der häufigen Verwendung hierfür am meisten geeignet. Es geht von einem Hauptgeneffekt aus, der durch einen biallelischen QTL definiert wird, und berücksichtigt zusätzlich einen Umwelteffekt, der als Familieneffekt für alle Familienmitglieder gleich ist, sowie einen Fehlerterm.

Der Anteil des Hauptgeneffektes an der gesamten Varianz des quantitativen Merkmales wurde auf 20 % festgelegt (Heritabiltät von 0,2 im weiteren Sinne). Dieser Wert ist realistisch, wie zahlreiche Schätzungen im Rahmen von Segregationsanalysen nahelegen. Hier seien nur zwei Beispiele aus der praktischen Arbeit des Autors genannt. Im Rahmen einer genetischen Kartierung der humanen Onchozerkose und einer weiteren Studie zur genetischen Kartierung der milden Malaria haben der Autor und seine Kollegen eine genortspezifische Heritabilität von 20 respektive 38% geschätzt (Timmann et al., 2007; Timmann et al., 2008).

Die Wahl der Familienstrukturen und der drei Studiendesigns wurde bereits in der Einleitung und Material und Methoden ausführlich dargestellt. Sie wird hier deshalb nicht nochmals erläutert (siehe S. 2 und S. 20f). Die Stichprobengröße pro Datensatz wurde mit 100 Kernfamilien mit zwei Geschwistern bzw. 300 Kernfamilien mit zwei bis fünf Geschwistern (mittlere Anzahl der Geschwister lag bei 3,04) ebenfalls auf eine in Studien gebräuchliche Größe festgelegt. Die bereits erwähnte Studie zur genetischen Kartierung der milden Malaria erfasste z.B. 108 Familien mit

einer mittleren Anzahl von 3,54 Geschwistern. In dem im Rahmen dieser Arbeit erneut analysierten COAG Perth Datensatz wurden 98 Familien mit zwei bis vier Geschwister pro Familie untersucht.
Als Marker für die genetische Kartierung wurde ein short-tandem-repeat Marker gewählt, dessen Informativität der eines typischen Markers entspricht (Heterozygosität von 90%). Seit einiger Zeit werden gehäuft Einzelnukleotid-Polymorphismus-Marker-Arrays sogenannte SNP-Arrays verwendet. Typischerweise ist die Informativität in der Multipoint-Analyse noch einmal ein wenig höher – z.B. bei der Verwendung des Affymetrix GeneChip Human Mapping 10K v2 Array in der bereits erwähnten genetischen Kartierung der milden Malaria lag der mittlere Information Content bei über 95 % (Timmann et al., 2007). Dieser Unterschied ist gering und sollte deshalb nur einen geringen Einfluss auf die Ergebnisse dieser Arbeit haben.
Die zuvor beschriebenen Modellannahmen wurden z.T. in gleicher oder ähnlicher Form in anderen Simulationsstudien verwendet. Diese Studie unterscheidet sich jedoch von ihren Vorgängern in folgender Hinsicht:
a) Während in anderen Studien der Effekt der Verletzung der Normalverteilungsannahmen durch eine simple Transformation des gesamten Phänotyps untersucht wurde (Allison et al., 1999; Allison et al., 2000; Sham et al., 2002; Cuenco et al., 2003; Szatkiewicz et al., 2003), ist im Rahmen dieser Arbeit ein anderer Ansatz gewählt worden. Hier wurde der Fehlerterm des Falconer und Mackay Modells zum einen aus einer Normalverteilung und zum anderen aus einer logarithmischen Normalverteilung simuliert. Dies erlaubt es, den Hauptgen- und Familieneffekt in seiner ursprünglichen Form zu erhalten, während der resultierende Phänotyp eine starke Schiefe und Wölbung aufweißt. Da die Modelle sich lediglich in der Verteilungsform des Fehlerterms unterscheiden, ist es damit auch möglich, einen direkten Powervergleich innerhalb der Verfahren zwischen diesen beiden Modellen anzustellen und damit eine Aussage über die Robustheit der Power der Verfahren zu treffen, wie im Abschnitt Zielsetzung (S. 5) gefordert.
b) Die meisten Simulationsstudien verwenden 10.000 Simulationen unter der Nullhypothese, um den empirischen Typ I-Fehler zu schätzen. Bei einem nominalen Typ I-Fehler von 0,05 ist dann der 95% Konfidenzintervall des empirischen Typ I-Fehlers ungefähr 0,01. Der Konfidenzintervall steigt bei niedrigeren nominalen Typ I-Fehlern stark an. Deshalb können bei niedrigeren nominalen Typ I-Fehlern die Wirkungen auf den empirischen Typ I-Fehler nur noch mit einer geringen Sicherheit beurteilt werden. In genomweiten Studien sind jedoch Aufgrund des multiplen Testproblems Typ I-Fehler von 0,001 und niedriger als Signifikanzkriterium zu wählen und deshalb von besonderem Interesse. Deshalb wurden hier 100.000 Simulationen unter der Nullhypothese für jedes Szenario gewählt – eine höhere Anzahl wäre mit aktuellen Computern für den gesamten Studienumfang in absehbarer Zeit praktisch nicht berechenbar gewesen.
Für diese Arbeit wurden acht in der Literatur beschriebene Verfahren gewählt, die derzeit häufig zur Kartierung quantitativer Merkmale in Kernfamilien verwendet werden. Die häufige Verwen-

dung der Verfahren gründet u.a. darauf, dass diese Verfahren in frei verfügbaren Softwarepaketen implementiert sind. Diese Verfahren werden daher in der Literatur und Übersichtsarbeiten häufig erwähnt.

## 5.2 Softwarepaket Sibsim

Mit dem Softwarepaket `Sibsim` steht das erforderliche Werkzeug zur Verfügung, welches die Simulation von quantitativen Phänotypen als auch Genotypen in einer sehr flexiblen Weise ermöglicht, wie sie zur Bestimmung der Güteeigenschaften von kopplungsanalytischen Verfahren und zur Validierung der Implementationen von neuen Kartierungsverfahren in Softwarepaketen benötigt werden (Franke et al., 2006). Die Veröffentlichung unter der Open-Source Lizenz GPL ermöglicht die freie Verwendung und Weiterentwicklung durch jedermann.

In der externen Validierung wurden alle Simulationsparameter für den speziell beabsichtigten Gebrauch geprüft und dokumentiert. Da sich keine Hinweise auf Fehler oder andere Auffälligkeiten zeigten, ist deshalb die Schlussfolgerung zulässig, dass die Simulationen im Rahmen dieser Arbeit den geforderten Simulationsparametern entsprechen.

Die Anwendung von `Sibsim` ist aber nicht auf diese Anwendung beschränkt. Darüber hinaus kann `Sibsim` auch für andere Fragestellungen wie z.B. die Ermittlung empirischer *p*-Werte bei der Analyse realer Datensätze eingesetzt werden. Im Rahmen einer genomweiten Kopplungsanalyse zur milden Malaria wurde `Sibsim` von Kollegen und dem Verfasser verwendet, um die notwendigen Permutationen zur Ermittlung der empirischen *p*-Werte der verwendeten kopplungsanalytischen Verfahren zu simulieren (Timmann et al., 2007). Hierzu wurden 100.000 mal ein genetischer Marker mit einer vergleichbaren Informativität, aber gleicher Familienstruktur simuliert, jeweils mit den ursprünglichen Phänotypen verknüpft und auf Basis dieser Simulationen empirische *p*-Werte ermittelt (für Details siehe Abschnitt 4).

## 5.3 Gütevergleich der kopplungsanalytischen Verfahren

Da die Ergebnisse dieser Arbeit nur für die Modellbedingungen dieser Studie generalisiert werden können, wurden wie bereits zu Beginn dieses Kapitels beschrieben die Modellbedingungen so realistisch wie möglich gewählt. Jedoch könnte gerade eine andere Wahl der Abweichungen von der Normalverteilung einen deutlichen Einfluss auf die Ergebnisse haben. Nachfolgend werden jetzt die ermittelten Güteeigenschaften der einzelnen Verfahren diskutiert.

Die im Vergleich niedrige Power des HASEMAN-ELSTON Verfahrens speziell im Vergleich zu MERLIN-REGRESS und insbesondere in selektierten Stichproben in Verbindung mit dem starken Einfluss unter Verletzung der Normalverteilungsannahmen auf den Typ I-Fehler macht es wohl unnötig, dieses Verfahren weiter zu verwenden.

Das REVIDIERTE HASEMAN-ELSTON Verfahren zeigt einen korrekten Typ I-Fehler bei unabhängigen Geschwisterschaften mit einer leichten Tendenz, unter Verletzung der Normalverteilungsannahmen konservativ zu werden. Leider ist der Typ I-Fehler bei abhängigen Geschwisterschaften deutlich zu liberal. Ein weiteres Problem ist die unter Normalverteilungsannahmen geringe Power im Vergleich zu MERLIN-REGRESS. Eine generelle uneingeschränkte Empfehlung zur Verwendung des REVIDIERTEN HASEMAN-ELSTON Verfahrens lässt sich deshalb aus den Ergebnissen dieser Arbeit nicht ableiten.

MERLIN-QTL mit der Whittemore und Halpern Teststatistik zeigte generell einen deutlich zu konservativen Typ I-Fehler. Kong und Cox (1997) geben hierfür als mögliche Erklärung an, dass für die Whittemore und Halpern Teststatistik ein nicht perfekt informativer genetischer Marker in einem zu konservativen Typ I-Fehler resultieren kann. Der Whittemore und Halpern Teststatistik Ansatz sollte deshalb nicht verwendet werden. Die anderen beiden Allele-sharing Verfahren MERLIN-QTL mit der Kong und Cox Teststatistik und das MAXIMUM LIKELIHOOD BINOMIAL Verfahren zeigten einen korrekten und robusten Typ I-Fehler unter allen Modellen, aber die Power ist mit Ausnahmen für die Studiendesigns unter einfacher und doppelter Selektion unter Verletzung der Normalverteilungsannahmen gering.

Ähnliches gilt für den WILCOXON-RANGSUMMENTEST. Jedoch scheint die Power hier besser als bei den Allele-sharing Verfahren zu sein. Die bemerkenswert höhere Power beim additiven Model unter Verletzung der Normalverteilungsannahmen und zufälliger Selektion im Vergleich zum dominanten und rezessiven Modell kann durch die Wahl der IBD-Zentrierungsfunktion in der Teststatistik wie sie von `Genehunter` verwendet wird erklärt werden. Kruglyak und Lander empfehlen die von `Genehunter` verwendete Zentrierung speziell für additive genetische Modelle (Kruglyak und Lander, 1995a; Kruglyak und Lander, 1995b).

Die Power der Allele-sharing Verfahren sowie der WILCOXON-RANGSUMMENTEST stiegen unter Verletzung der Normalverteilungsannahmen gegenüber den gleichen Simulationen unter Normalverteilungsannahmen. Mit einigen Ausnahmen war dieser Effekt genau umgekehrt für alle anderen untersuchten Verfahren. Der Grund hierfür könnte das hohe dritte Moment der Phänotyp-Verteilung unter Verletzung der Normalverteilungsannahmen sein. Die Phänotypverteilung an jedem der drei Genotypen des biallelischem QTL ist ebenfalls rechts schief, jedoch mit den gleichen Erwartungswerten wie unter Normalverteilungsannahmen (siehe Kapitel 3, Abschnitt 1.1, S. 18f). Die Allele-sharing Verfahren sowie der WILCOXON-RANGSUMMENTEST scheinen dieses besser verwenden zu können.

Zusammenfassend lässt sich jedoch sagen, dass Aufgrund der niedrigen Power unter Normalverteilungsannahmen die Verwendung der Allele-sharing Verfahren sowie des WILCOXON-RANGSUMMENTEST nur unter bestimmten Umständen zu empfehlen ist. In speziellen Studien-

designs, z.B. wenn der Phänotyp ordinal verteilt ist, könnten diese Verfahren – insbesondere der WILCOXON-RANGSUMMENTEST – aber die Methode der Wahl sein.

Auf die Problematik der Verwendung des voll parametrisierten MODELLBASIERTEN Verfahrens ist in dieser Arbeit mehrfach hingewiesen worden. Sie wird deshalb an dieser Stelle nicht erneut diskutiert (siehe Kapitel 2, Abschnitt 8, S. 17). Jedoch sollte an dieser Stelle noch erwähnt werden, dass der konservative Typ I-Fehler bereits von Rao und Kollegen (1978) für eine Vielzahl von empirischen Studien durch Vergleich der beobachteten zu den nominalen Typ I-Fehlern gezeigt wurde.

Die Ergebnisse zu den VARIANZKOMPONENTENMODELLEN zeigen, dass auch unter Normalverteilungsannahmen und bei einem realistischen Stichprobenumfang ein substantiell erhöhter Typ I-Fehler resultiert. Zwei Gründe können dieses unerwartete Ergebnis erklären: Ferreira hat darauf hingewiesen, dass die zu schätzende Varianz-Kovarianzmatrix sechs Parameter enthält (Ferreira, 2004). Wenn dieses Verfahren jedoch auf Kernfamilien angewendet wird, sind diese Parameter unteridentifiziert. Zusätzliche Restriktionen sind notwendig, um diese Parameter schätzen zu können. Darüber hinaus sind Standardfehler dieser Varianzen Momente vierter Ordnung. Es ist allgemein bekannt, dass Statistiken, die auf Momenten vierter Ordnung beruhen für relativ geringe Stichprobengrößen, die für reale Studien typisch sind, instabil sind (siehe z.B. Bentler und Dudgeon, 1996).

Einige Empfehlungen werden in der Literatur gegeben, um die Robustheit des Typ I-Fehlers zu verbessern. Eine gängige Empfehlung ist z.B., den Phänotyp in der Hoffnung, er sei dann multivariat normalverteilt, mit einer Funktion so zu transformieren, dass er scheinbar einer Normalverteilung folgt. Leider gibt es zum einen keine Garantie, dass solch eine Funktion wirklich existiert. Zum anderen lässt sich aus der scheinbaren Normalverteilung des Phänotypen nicht folgern, dass der Phänotyp wirklich einer multivariaten Normalverteilung folgt. Als Beispiel für die Problematik brauchen wir nur die Phänotypen unter Verletzung der Normalverteilungsannahmen betrachten, wie sie im Rahmen dieser Arbeit verwendet wurden. Es existiert keine Funktion, mit der sich diese Phänotypen in eine multivariate Normalverteilung transformieren lassen.

Als zweite Empfehlung wird immer wieder die Verwendung robuster Schätzer genannt (Blangero et al., 2000). Jedoch kann diese zu einer Reduktion der Power führen. Als dritte Möglichkeit, insbesondere für die Verwendung auf selektierte Stichproben, könnten die Selektions-Wahrscheinlichkeiten in die Kalkulation der Likelihood mit einfließen. Auch dies kann aber zu einer Reduktion der Power führen und würde die Problematik des substantiell erhöhten Typ I-Fehlers unter Normalverteilungsannahmen und bei einem realistischen Stichprobenumfang nicht lösen.

Wenn, wie von einigen anderen Autoren (Sham et al., 2002; Yu et al., 2004) und den Ergebnissen dieser Arbeit ebenfalls unterstützt, die VARIANZKOMPONENTENMODELLEN und MERLIN-REGRESS eine ähnliche Power zeigen, dann ist es aus den zuvor erläuterten Gründen nicht mehr notwendig,

die VARIANZKOMPONENTENMODELLE weiterhin zu verwenden. Im Vergleich zu den VARIANZKOMPONENTENMODELLEN zeigt MERLIN-REGRESS sehr robuste Typ I-Fehler. Diese werden auch nicht von selektierten Stichproben beeinflusst. Darüber hinaus wurde im Rahmen dieser Arbeit gezeigt, dass der Typ I-Fehler nicht durch eine Ein-Parameter-Misspezifikation beeinflusst wird. Dies gilt auch für die komplexere Situation einer selektierten Stichprobe in Kombination mit einer Verletzung der Normalverteilungsannahmen. Lediglich der leicht liberale Typ I-Fehler in der Situation der abhängigen Geschwisterschaften - wie er in dieser Arbeit gezeigt worden ist – ist von Nachteil.

Aus der generell hohen und robusten Power in Verbindung mit dem robusten Typ I-Fehler lässt sich eine generelle Empfehlung zur Verwendung dieses Verfahrens für die Kartierung von quantitativen Merkmalen in Kernfamilien mit unabhängigen Geschwisterschaften ableiten. Lediglich der leicht erhöhte Typ I-Fehler in der Situation der abhängigen Geschwisterschaften realistischer Größe ist ein leichter Nachteil. Eine generelle Empfehlung der Anwendung auf Kernfamilien mit abhängigen Geschwisterschaften beliebiger Größe, sowie erweiterten Familienstammbäumen lässt sich jedoch nicht ableiten. So haben z.B. Huang und Kollegen (2007) auf dem Genetic Analysis Workshop 15 einen deutlich liberalen Typ I-Fehler bei der Anwendung des Verfahrens auf erweiterte Stammbäume identifiziert.

Feingold (2002, S. 220) kommt in einer Übersichtsarbeit aufgrund der von Sham und Kollegen in der Originalveröffentlichung gezeigten Eigenschaften zu dem Schluss, dass MERLIN-REGRESS viele der erwünschten Eigenschaften aufweist („Sham et al. provide extensive sumulation results, which suggest that the method does indeed have many of the properties we would like.“). Die Ergebnisse dieser Arbeit können diese Aussage bis auf die Einschränkung des liberalen Typ I-Fehlers in der Situation der abhängigen Geschwisterschaften nur unterstützen.

## 5.4 Analyse des COAG Perth Datensatzes

Die Ergebnisse der Analyse decken sich im Wesentlichen mit den Ergebnissen der Powervergleiche der verschiedenen Verfahren unter zufälliger Selektion, wie sie im Kapitel 4, Abschnitt 4 vorgestellt worden sind. Während die VARIANZKOMPONENTENMODELLEN und MERLIN-REGRESS die niedrigsten *p*-Werte zeigen, sind die *p*-Werte des HASEMAN-ELSTON Verfahrens sowie der Allele-sharing Verfahren im Vergleich zu den anderen sehr hoch.

Noch ein zweiter wichtiger Punkt lässt sich an der Analyse des COAG Perth Datensatzes erkennen. Die *p*-Werte für den gegen Alter, Geschlecht und Antigen spezifischen IgE Titer korrigierten logarithmierten IgE Titer sind kleiner, als für den unkorrigierten logarithmierten IgE Titer. Unter der Annahme, dass die untersuchte genetische Region wirklich mit dem QTL gekoppelt ist, sollte die Korrektur des Phänotyps - z.B. über geeignete Regressionsmodelle - den *p*-Wert der geneti-

schen Kartierungsanalyse erniedrigen, da der Varianzanteil des genetischen Effektes vergrößert wird. Auch für die bereits zuvor erwähnten Studien zur humanen Onchozerkose und milden Malaria wurden die Phänotypen vor der Verwendung gegen Kovariaten mit statistisch signifikantem Einfluss korrigiert (Timmann et al., 2007; Timmann et al., 2008).

Noch ein dritter wichtiger Punkt, nämlich die Ermittlung empirischer *p*-Werte, sollte an dieser Stelle erwähnt werden. Bei der Analyse von Datensätzen bietet es sich immer an, empirische *p*-Werte zu ermitteln und diese zu verwenden. In den Softwarepaketen `Merlin` und `S.A.G.E.` sind diese Möglichkeiten bereits implementiert. Prinzipiell beruhen diese Verfahren alle auf der gleichen Idee: Durch Permutation einer Variablen wird die Kopplung des Phänotyps und Genotyps aufgehoben. Danach wird die Teststatistik mit diesem permutierten Datensatz errechnet. Dieses wird dann z.B. 100.000 mal wiederholt und ergibt dann die empirische Verteilung der Teststatistik unter der Nullhypothese, die dann wiederum verwendet wird, um den empirischen *p*-Wert zu ermitteln (siehe z.B. auch Ziegler und König, 2010, Kapitel 9).

`S.A.G.E.` permutiert hierzu die Phänotypen zwischen Familien gleicher Größe als ganzes. Dieses Verfahren ist einer Permutation der Phänotypen innerhalb der Familie vorzuziehen, da mögliche Residualkorrelationen unter den Geschwistern erhalten bleiben. `Merlin` geht einen anderen Weg und permutiert direkt die geschätzten IBD Vererbungsvektoren. Ein dritter Weg mit dem Softwarepaket `Sibsim`, der keinen Eingriff in den Quellcode der Software notwendig macht, ist bereits im Abschnitt zuvor beschrieben worden und lässt sich deshalb prinzipiell auf alle Verfahren anwenden.

## 5.5 Ausblick

Genomweite Assoziationsstudien stehen derzeit im Mittelpunkt des Interesses. Jedoch bleiben Kopplungsanalysen, wie von Darpoux und Elston (2007) in einem Übersichtsartikel herausgestellt, ein wichtiges Werkzeug zur genetischen Kartierung von Merkmalen.

Da eine Reihe wichtiger Erkrankungen und Merkmaler eine quantitative Ausprägung zeigen und sich deshalb im Allgemeinen präziser durch eine quantitative Definition des Merkmales beschreiben lassen, wird auch in Zukunft die genetische Kartierung quantitativer Merkmale eine wichtige Rolle in der genetischen Epidemiologie spielen. Die Ergebnisse dieser Arbeit helfen bei der Auswahl des Verfahrens und Studiendesigns. Darüber hinaus steht mit der Software `Sibsim` ein universelles flexibles Werkzeug zur Verfügung, um auch in der Zukunft die Güteeigenschaften von Verfahren unter einer Vielzahl von verschiedenen Bedingungen zu vergleichen und mit entsprechenden Anpassungen im Quellcode auch anderen Fragestellungen zu bearbeiten.

Zur Durchführbarkeit dieses Gütevergleiches war es notwendig, die analysierten Szenarien zu beschränken. Im Rahmen dieser Arbeit wurden deshalb Effekte wie genetische Heterogenität,

Genomisches Imprinting, Gen-Umwelt Interaktionen oder Fehler in den Daten nicht untersucht. Jedoch eröffnet auch hierfür die Simulationssoftware `Sibsim` mit entsprechenden Anpassungen im Quellcode eine effiziente Möglichkeit für entsprechende Studien in der Zukunft.

Abschließend soll noch einmal auf die Analyse der COAG Perth Studie eingegangen werden: In dem Datensatz wurden nur zwei genetische Marker in einer Kandidatenregion untersucht. Bei entsprechenden genomweiten Studien insbesondere bei der Verwendung von SNP-Arrays mit mehreren 100.000 SNPs nimmt der Bedarf an bioinformatischer Unterstützung große Ausmaße an. Dieses verdeutlicht exemplarisch die steigende Notwendigkeit der interdisziplinären Zusammenarbeit von Experten aus verschiedenen Disziplinen wie Informatik, Statistik, Genetik, Medizin und Biochemie besonders in dem Forschungsbereich der genetischen Epidemiologie.

# 6 Zusammenfassung

In der Praxis werden zur Aufklärung komplexer genetischer Erkrankungen zunehmend kopplungsanalytische Methoden für quantitative Phänotypen unter Verwendung von Kernfamilien mit zwei oder mehr Geschwistern eingesetzt. In den letzten Jahren wurde eine Vielzahl neuer Verfahren für diese Fragestellungen entwickelt, jedoch ist bisher weitgehend ungeklärt, wie sich die Güte dieser Verfahren im direkten Vergleich zueinander verhält. Daher wurde im Rahmen dieser Arbeit in einer Monte-Carlo Simulationsstudie die Güteeigenschaften von insgesamt acht in der Praxis häufig eingesetzten Verfahren unter verschiedenen Modellen und Studiendesigns verglichen. Diese Verfahren wurden unter drei genetischen Modellen (dominant, additiv, rezessiv) drei Studiendesigns (ohne Selektion, mit einfacher Selektion und doppelter Selektion) und zwei Familienstrukturen (Kernfamilien mit einem Geschwisterpaar sowie Kernfamilien mit einer variierenden Anzahl von zwei bis fünf Geschwistern) untersucht. Zusätzlich wurde der Effekt bei Abweichung von der Normalverteilung untersucht.

In einem ersten Schritt wurde hierzu zunächst eine Simulationssoftware erstellt und extern validiert (`Sibsim`), anhand derer die Datensätze für die 36 Szenarien simuliert wurden. Für den Robustheitsvergleich wurden dann für jedes Szenario 100.000 Simulationen unter der Nullhypothese (keine Kopplung) und zum Powervergleich 1.000 Simulationen unter der Alternativhypothese (vollständige Kopplung) erstellt. Der Robustheitsvergleich wurde durch Vergleich der Abweichungen zwischen den empirisch ermittelten Typ I-Fehleranteilen und dem nominalen Typ I-Fehler auf verschiedenen Testniveaus durchgeführt. Die hohe Anzahl der Simulationen unter der Nullhypothese ermöglicht dann auf Basis empirisch ermittelter Grenzwerte einen empirischen Powervergleich unter der Alternativhypothese.

Die Anwendung der verschiedenen im Rahmen dieser Arbeit verwendeten Verfahren wurde sodann an dem Datensatz „Consortium on Asthma Genetics: Perth study“ illustriert (Palmer et al., 1998; Palmer et al., 2001), wobei besonderer Wert auf die Illustrierung der praktischen Vorgehensweise gelegt wurde.

Im Rahmen dieser Arbeit wurden zum ersten Mal die Güteeigenschaften für eine solche Vielzahl von kopplungsanalytischen Verfahren zur Kartierung quantitativer Merkmale unter einer Vielzahl verschiedener realistischer Annahmen und Studiendesigns in einem direkten Vergleich ermittelt, verglichen und ausführlich diskutiert. Die Ergebnisse dieser Arbeit können als wertvolle Quelle bei der Auswahl des Verfahrens und Studiendesigns für Kartierungsstudien von quantitativen Merkmalen dienen. Darüber hinaus steht mit der Software `Sibsim` ein universelles flexibles Werkzeug zur Verfügung, um auch in der Zukunft die Güteeigenschaften von Verfahren unter einer Vielzahl von verschiedenen Bedingungen zu vergleichen und mit entsprechenden Anpassungen im Quellcode auch anderen Fragestellungen zu bearbeiten.

# 7 Softwarepakete und Literaturverzeichnis

## 7.1 Softwarepakete

Die folgende Liste enthält die im Rahmen dieser Arbeit verwendeten Softwarepakete. Angegeben sind soweit verfügbar das Erscheinungsjahr der verwendeten Version inkl. der Versionsnummer, Zitation der zugehörigen Veröffentlichung sowie die Quelle im Internet (Tag des letzten Zugriffs: 01.06.2009).

Die meisten der hier aufgeführten Programme sind beim Erscheinen dieser Arbeit bereits in neueren Versionen verfügbar.

`Abi2Link` (2003) Version 1.0

http://www.imbs-luebeck.de/imbs/de/software

`Genehunter` (2003) Version 2.1_r4

(Kruglyak et al., 1996)

http://www.broadinstitute.org/ftp/distribution/software/genehunter/

`Linkage` (1995) Version 5.1

(Lathrop et al., 1984)

ftp://linkage.rockefeller.edu/software/linkage

`Mega2` (2003) A Manipulation Environment for Genetic Analyses, Version 2.5

(Mukhopadhyay et al., 2005)

http://watson.hgen.pitt.edu/mega2.html

`Merlin` (2004) Rapid analysis of dense genetic maps using sparse gene flow trees,

Version 0.10.2

(Abecasis et al., 2002; Sham et al., 2002)

http://www.sph.umich.edu/csg/abecasis/Merlin/download/

`Mlbgh` (1998) Maximum likelihood binomial GENEHUNTER, Version 1.0

(Abel und Müller-Myhsok, 1998; Alcaïs und Abel, 1999)

http://genamics.com/software/downloads/mlbgh-1.0.tar.Z

`Pedcheck` (1998), Version 1.00

(O'Connell und Weeks, 1998)

http://watson.hgen.pitt.edu/register/soft_doc.html

`R` (2005) Free software environment for statistical computing and graphics, Version 2.1.0

http://www.R-project.org

`S.A.G.E.` (2003) Statistical Analysis for Genetic Epidemiology, Version 4.3

http://darwin.cwru.edu/sage/

`Sibsim` (2003) Version 1.02

(Franke et al., 2006)

http://www.imbs-luebeck.de/imbs/de/software

`Solar` (2003) Sequential Oligogenic Linkage Analysis Routines, Version 1.7.4

(Almasy und Blangero, 1998)

http://solar.sfbrgenetics.org/download.html

## 7.2 Literaturverzeichnis

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002): Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97-101

Abel L, Alcaïs A, Mallet A (1998): Comparison of four sib-pair linkage methods for analyzing sibships with more than two affecteds: interest of the binomial maximum likelihood approach. Genet Epidemiol 15:371-390

Abel L, Müller-Myhsok B (1998): Robustness and power of the maximum-likelihood-binomial and maximum-likelihood-score methods, in multipoint linkage analysis of affected-sibship data. Am J Hum Genet 63:638-647

Alcaïs A, Abel L (1999): Maximum-likelihood-binomial method for genetic model-free linkage analysis of quantitative traits in sibships. Genet Epidemiol 17:102-117

Allison DB, Fernández JR, Heo M, Beasley TM (2000): Testing the robustness of the new Haseman-Elston quantitative-trait loci-mapping procedure. Am J Hum Genet 67:249-252

Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J (1999): Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet 65:531-544

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

Amos CI (1994): Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535-543

Amos CI, Elston RC (1989): Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol 6:349-360

Amos CI, Elston RC, Bonney GE, Keats BJ, Berenson GS (1990): A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. Am J Hum Genet 47:247-254

Amos CI, Zhu DK, Boerwinkle E (1996): Assessing genetic linkage and association with robust components of variance approaches. Ann Hum Genet 60 (Pt 2):143-160

Bentler PM, Dudgeon P (1996): Covariance structure analysis: statistical practice, theory, and directions. Annu Rev Psychol 47:563-592

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

Blangero J (2004): Localization and identification of human quantitative trait loci: king harvest has surely come. Curr Opin Genet Dev 14:233-240

Blangero J, Williams JT, Almasy L (2000): Robust LOD scores for variance component-based linkage analysis. Genet Epidemiol 19 Suppl 1:S8-14

Blangero J, Williams JT, Almasy L (2001): Variance component methods for detecting complex trait loci. Adv Genet 42:151-181.

Carey G, Williamson J (1991): Linkage analysis of quantitative traits: increased power by using selected samples. Am J Hum Genet 49:786-796.

Clerget-Darpoux F, Elston RC (2007): Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. Hum Hered 64:91-96

Commenges D (1994): Robust genetic linkage analysis based on a score test of homogeneity: the weighted pairwise correlation statistic. Genet Epidemiol 11:189-200

Cuenco KT, Szatkiewicz JP, Feingold E (2003): Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. Am J Hum Genet 73:863-873

Dempster AP, Laird NM, Rubin DB (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion). JRSS B 39:1-38

Dina C, Nemanov L, Gritsenko I, Rosolio N, Osher Y, Heresco-Levy U, Sariashvilli E, Bachner-Melman R, Zohar AH, Benjamin J, Belmaker RH, Ebstein RP (2005): Fine mapping of a region on chromosome 8p gives evidence for a QTL contributing to individual differences in an anxiety-related personality trait: TPQ harm avoidance. Am J Med Genet B Neuropsychiatr Genet 132:104-108

Dolan CV, Boomsma DI (1998): Optimal selection of sib pairs from random samples for linkage analysis of a QTL using the EDAC test. Behav Genet 28:197-206

Drigalenko E (1998): How sib pairs reveal linkage. Am J Hum Genet 63:1242-1245

Duggirala R, Williams JT, Williams-Blangero S, Blangero J (1997): A variance component approach to dichotomous trait linkage analysis using a threshold model. Genet Epidemiol 14:987-992

Elston RC (1998): Linkage and association. Genet Epidemiol 15:565-576

Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000): Haseman and Elston revisited. Genet Epidemiol 19:1-17

Falconer DS, Mackay TFC (1996): Introduction to quantitative genetics. 4. Auflage, Longman, Essex, England

Feingold E (2002): Regression-based quantitative-trait-locus mapping in the 21st century. Am J Hum Genet 71:217-222.

Ferreira MA (2004): Linkage analysis: principles and methods for the analysis of human quantitative traits. Twin Res 7:513-530

Fisher RA (1918): The correlation between relatives on the supposition of mendelian inheritance. Trans Roy Soc 52:399-433

Forrest WF (2001): Weighting improves the "new Haseman-Elston" method. Hum Hered 52:47-54

Franke D, Kleensang A, Ziegler A (2006): SIBSIM - quantitative phenotype simulation in extended pedigrees. GMS Med Inform Biom Epidemiol 2:Doc4

Fulker DW, Cardon LR, DeFries JC, Kimberling WJ, Pennington BF, Smith SD (1991): Multiple regression analysis of sib-pair data on reading to detect quantitative trait loci. Read Writ Interdisciplinary J 3:299-313

Fulker DW, Cherny SS (1996): An improved multipoint sib-pair analysis of quantitative traits. Behav Genet 26:527-532

Gillham NW (2001): Evolution by jumps: Francis Galton and William Bateson and the mechanism of evolutionary change. Genetics 159:1383-1392

Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L, Mant R, Newton P, Rooke K, Roques P, Talbot C, Pericak-Vance M, Roses A, Williamson R, Rossor M, Owen M, Hardy J (1991): Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349:704-706

Gu C, Todorov A, Rao DC (1996): Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. Genet Epidemiol 13:513-533

Haseman JK, Elston RC (1972): The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3-19

Huang S, Ballard D, Zhao H (2007): The role of heritability in mapping expression quantitative trait loci. BMC Proc 1 Suppl 1:S86

Knoblauch H, Müller-Myhsok B, Busjahn A, Ben Avi L, Bähring S, Baron H, Heath SC, Uhlmann R, Faulhaber HD, Shpitzen S, Aydin A, Reshef A, Rosenthal M, Eliav O, Mühl A, Lowe A, Schurr D, Harats D, Jeschke E, Friedlander Y, Schuster H, Luft FC, Leitersdorf E (2000): A cholesterol-lowering gene maps to chromosome 13q. Am J Hum Genet 66:157-166

Kong A, Cox NJ (1997): Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179-1188

Korstanje R, Paigen B (2002): From QTL to gene: the harvest begins. Nat Genet 31:235-236

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996): Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347-1363.

Kruglyak L, Lander ES (1995a): Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439-454.

Kruglyak L, Lander ES (1995b): A nonparametric approach for mapping quantitative trait loci. Genetics 139:1421-1428.

Lathrop GM, Lalouel JM, Julier C, Ott J (1984): Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci U S A 81:3443-3446

Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE (2005): Mega2: data-handling for facilitating genetic linkage and association analyses. Bioinformatics 21:2556-2557

O'Connell JR, Weeks DE (1998): PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259-266

Palmer LJ, Cookson WO, Deichmann KA, Holloway JW, Laitinen T (2001): Single region linkage analyses of asthma: description of data sets. Genet Epidemiol 21 Suppl 1:S9-15

Palmer LJ, Daniels SE, Rye PJ, Gibson NA, Tay GK, Cookson WO, Goldblatt J, Burton PR, LeSouef PN (1998): Linkage of chromosome 5q and 11q gene markers to asthma-associated quantitative traits in Australian children. Am J Respir Crit Care Med 158:1825-1830

Palmer LJ, Jacobs KB, Elston RC (2000): Haseman and Elston revisited: the effects of ascertainment and residual familial correlations on power to detect linkage. Genet Epidemiol 19:456-460

Perola M, Sammalisto S, Hiekkalinna T, Martin NG, Visscher PM, Montgomery GW, Benyamin B, Harris JR, Boomsma D, Willemsen G, Hottenga JJ, Christensen K, Kyvik KO, Sorensen TI, Pedersen NL, Magnusson PK, Spector TD, Widen E, Silventoinen K, Kaprio J, Palotie A, Peltonen L (2007): Combined genome scans for body stature in 6,602 european twins: evidence for common caucasian loci. PLoS Genet 3:e97

Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, Perusse L, Bouchard C (2006): The human obesity gene map: the 2005 update. Obesity (Silver Spring) 14:529-644

Rao DC, Keats BJ, Morton NE, Yee S, Lew R (1978): Variability of human linkage data. Am J Hum Genet 30:516-529

Risch N, Zhang H (1995): Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584-1589

Sham PC, Purcell S (2001): Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. Am J Hum Genet 68:1527-1532.

Sham PC, Purcell S, Cherny SS, Abecasis GR (2002): Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71:238-253.

Single RM, Finch SJ (1995): Gain in efficiency from using generalized least squares in the Haseman-Elston test. Genet Epidemiol 12:889-894

Speer MC, Terwilliger JD, Ott J (1995): Data simulation for GAW9 problems 1 and 2. Genet Epidemiol 12:561-564

Streeten EA, McBride DJ, Pollin TI, Ryan K, Shapiro J, Ott S, Mitchell BD, Shuldiner AR, O'Connell JR (2006): Quantitative trait loci for BMD identified by autosome-wide linkage scan to chromosomes 7q and 21q in men from the Amish Family Osteoporosis Study. J Bone Miner Res 21:1433-1442

Szatkiewicz JP, K TC, Feingold E (2003): Recent advances in human quantitative-trait-locus mapping: comparison of methods for discordant sibling pairs. Am J Hum Genet 73:874-885

Terwilliger JD, Goring HH (2000): Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. Hum Biol 72:63-132

Terwilliger JD, Ott J (1994): Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore, USA

Timmann C, Evans JA, König IR, Kleensang A, Rüschendorf F, Lenzen J, Sievertsen J, Becker C, Enuameh Y, Kwakye KO, Opoku E, Browne EN, Ziegler A, Nürnberg P, Horstmann RD (2007): Genome-wide linkage analysis of malaria infection intensity and mild disease. PLoS Genet 3:e48

Timmann C, van der Kamp E, Kleensang A, König IR, Thye T, Büttner DW, Hamelmann C, Marfo Y, Vens M, Brattig N, Ziegler A, Horstmann RD (2008): Human genetic resistance to Onchocerca volvulus: evidence for linkage to chromosome 2p from an autosome-wide scan. J Infect Dis 198:427-433

Whittemore AS, Halpern J (1994): A class of tests for linkage using affected pedigree members. Biometrics 50:118-127

Williams JT, Blangero J (1999): Comparison of variance components and sibpair-based approaches to quantitative trait linkage analysis in unselected samples. Genet Epidemiol 16:113-134

Wright FA (1997): The phenotypic difference discards sib-pair QTL linkage information. Am J Hum Genet 60:740-742

Yu X, Knott SA, Visscher PM (2004): Theoretical and empirical power of regression and maximum-likelihood methods to map quantitative trait loci in general pedigrees. Am J Hum Genet 75:17-26

Ziegler A (2001): The new Haseman-Elston method and the weighted pairwise correlation statistic are variations on the same theme. Biometrical Journal 43:697-702

Ziegler A, König IR (2010): A statistical approach to genetic epidemiology: concepts and applications. 2. Auflage, Wiley-VCH, Weinheim

Zmuda JM, Sheu YT, Moffett SP (2006): The search for human osteoporosis genes. J Musculoskelet Neuronal Interact 6:3-15

# 8 Anhänge

## 8.1 Simulationsparameter Validierungs-Simulationen Sibsim

**Tabelle 22**: Verwendeten Simulationsparameter zur Validierung des Softwarepaketes `Sibsim`.

| Validierungs-Simulationen | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Startzufallszahl | 62634 | 7362 | 83615 | 30614 | 17350 | 97423 | 31065 | 67802 |
| Genetisches Modell | Dom | Add | Rez | Dom | Dom | Dom | Add | Rez |
| Varianz Hauptgeneffekt | 2,0 | 2,0 | 2,0 | 4,0 | 4,0 | 2,0 | 2,0 | 2,0 |
| Varianz Familieneffekt | 0 | 0 | 0 | 0 | 0 | 0,2 | 0,2 | 0,2 |
| Varianz Fehlerterm | 0,2 | 0,2 | 0,2 | 0,5 | 0,5 | 0 | 0 | 0 |
| Verteilung Fehlerterm | NV | NV | NV | NV | log NV | --- | --- | --- |
| Für alle Simulationen gleich: | | | | | | | | |
| Frequenz des hohen Allels | 0,5 | | | | | | | |
| Anzahl Familien | 600 | | | | | | | |
| Genetischer Marker | 10 Allele mit der Frequenz 10% | | | | | | | |

## 8.2 Startzufallszahlen für Monte-Carlo Simulationen

**Tabelle 23**: Startzufallszahlen, die zum Erstellen der Monte-Carlo Simulationen mit dem Programm `Sibsim` verwendet wurden.

| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Normalverteilungsannahmen: | | | | | | | | | | |
| Nullhypothese | unabhängig | 13092 | 71060 | 33132 | 80703 | 63003 | 53203 | 85197 | 59650 | 42067 |
| | abhängig | 44133 | 44167 | 74211 | 66986 | 70009 | 22965 | 54273 | 58064 | 53890 |
| Alternativhypothese | | 16408 | 81953 | 47498 | 00770 | 41548 | 64933 | 15178 | 20910 | 95588 |
| | | 21199 | 72163 | 81651 | 88048 | 32466 | 54955 | 46808 | 57167 | 57515 |
| Verletzung der Normalverteilungsannahmen: | | | | | | | | | | |
| Nullhypothese | | 45799 | 52390 | 22164 | 63282 | 42579 | 52667 | 86458 | 69155 | 67788 |
| | | 12426 | 56302 | 06116 | 14631 | 91178 | 85961 | 48853 | 81636 | 56344 |
| Alternativhypothese | | 49323 | 15059 | 85900 | 50993 | 44488 | 38646 | 23703 | 46829 | 06683 |
| | | 28225 | 34925 | 68335 | 83867 | 64361 | 22644 | 49044 | 75470 | 12293 |

## 8.3 Empirische Typ I-Fehler

### 8.3.1 Unter Normalverteilungsannahmen

**Tabelle 24**: Empirischer Typ I-Fehler [%] bei einem nominalen Fehlerniveau von 5% unter Normalverteilungsannahmen.

| | | Empirischer Typ I-Fehler unter Normalverteilungsannahmen | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Gh.HE.Trad | Unabhängig | 5,14 | 5,09 | 5,08 | 5,14 | 5,10 | 5,02 | 5,07 | 5,06 | 5,12 |
| Sage.HE | | 5,09 | 5,02 | 4,99 | 5,07 | 5,02 | 4,96 | 5,00 | 4,99 | 5,05 |
| rHE | | 4,91 | 4,99 | 5,05 | 5,15 | 5,08 | 4,90 | 4,92 | 4,89 | 5,01 |
| Merlin-Regress | | 5,03 | 5,00 | 5,11 | 5,18 | 5,08 | 4,84 | 5,04 | 4,97 | 5,06 |
| Gh.VC | | 6,14 | 5,44 | 6,10 | 5,18 | 4,04 | 2,86 | 7,53 | 7,38 | 7,52 |
| Solar.VC | | 6,10 | 5,40 | 6,07 | 5,12 | 4,05 | 2,84 | 7,52 | 7,35 | 7,47 |
| Npar | | 5,07 | 4,94 | 5,03 | 4,97 | 5,06 | 4,99 | 4,91 | 5,02 | 5,00 |
| Merlin.K&C | | 4,62 | 4,56 | 4,45 | 4,51 | 4,57 | 4,42 | 4,41 | 4,37 | 4,52 |
| Merlin.W&H | | 3,65 | 3,71 | 3,57 | 3,62 | 3,64 | 3,61 | 3,54 | 3,53 | 3,60 |
| Mlbqt.NV | | 5,21 | 5,24 | 5,13 | 5,17 | 5,19 | 5,09 | 4,91 | 4,90 | 5,05 |
| Mlbqt.Kat | | 5,14 | 5,29 | 5,13 | 5,15 | 5,17 | 5,07 | 4,94 | 4,91 | 5,01 |
| Linkage | | 4,92 | 1,96 | 4,01 | 5,24 | 4,32 | 5,16 | 5,14 | 4,89 | 5,24 |
| Gh.HE.Trad | Abhängig | 5,21 | 5,20 | 5,26 | 5,19 | 5,25 | 5,30 | 5,22 | 5,23 | 5,17 |
| Sage.HE | | 5,31 | 5,28 | 5,36 | 5,07 | 5,23 | 5,21 | 5,15 | 5,09 | 5,17 |
| rHE | | 5,97 | 5,90 | 5,90 | 5,68 | 5,52 | 5,71 | 5,27 | 5,25 | 5,18 |
| Merlin-Regress | | 5,47 | 5,45 | 5,40 | 5,30 | 5,40 | 5,51 | 5,27 | 5,25 | 5,24 |
| Gh.VC | | 5,98 | 5,28 | 5,93 | 5,38 | 4,91 | 5,18 | 5,78 | 5,32 | 5,76 |
| Solar.VC | | 5,94 | 5,20 | 5,89 | 5,35 | 4,88 | 5,16 | 5,69 | 5,28 | 5,72 |
| Npar | | 5,21 | 5,27 | 5,23 | 5,13 | 5,24 | 5,26 | 5,27 | 5,15 | 5,20 |
| Merlin.K&C | | 4,53 | 4,64 | 4,52 | 4,45 | 4,63 | 4,60 | 4,47 | 4,56 | 4,51 |
| Merlin.W&H | | 3,79 | 3,87 | 3,82 | 3,72 | 3,85 | 3,85 | 3,71 | 3,76 | 3,76 |
| Mlbqt.NV | | 5,03 | 5,02 | 4,94 | 4,85 | 4,97 | 5,09 | 4,89 | 4,86 | 4,90 |
| Mlbqt.Kat | | 4,95 | 5,07 | 4,95 | 4,84 | 4,98 | 5,05 | 4,91 | 4,90 | 4,92 |
| Linkage | | 4,62 | 1,98 | 3,84 | 5,04 | 3,69 | 4,91 | 5,05 | 3,60 | 4,90 |

**Tabelle 25**: Empirischer Typ I-Fehler [%] bei einem nominalen Fehlerniveau von 1% unter Normalverteilungsannahmen.

| | | Empirischer Typ I-Fehler unter Normalverteilungsannahmen | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Gh.HE.Trad | Unabhängig | 0,966 | 0,992 | 0,966 | 1,008 | 1,018 | 0,963 | 1,014 | 1,010 | 1,080 |
| Sage.HE | | 0,963 | 0,967 | 0,948 | 0,991 | 0,985 | 0,940 | 1,006 | 0,988 | 1,060 |
| rHE | | 0,990 | 0,998 | 1,009 | 1,024 | 1,044 | 0,932 | 1,021 | 0,989 | 1,018 |
| Merlin-Regress | | 0,958 | 0,970 | 0,963 | 1,029 | 1,014 | 0,961 | 0,994 | 1,007 | 0,988 |
| Gh.VC | | 1,430 | 1,145 | 1,417 | 0,675 | 0,393 | 0,170 | 2,073 | 2,032 | 2,167 |
| Solar.VC | | 1,414 | 1,142 | 1,417 | 0,678 | 0,391 | 0,173 | 2,052 | 2,027 | 2,157 |
| Npar | | 1,041 | 0,990 | 1,034 | 0,989 | 0,948 | 0,972 | 0,959 | 0,988 | 0,966 |
| Merlin.K&C | | 0,994 | 0,996 | 0,966 | 0,953 | 0,929 | 0,919 | 0,954 | 0,994 | 0,960 |
| Merlin.W&H | | 0,674 | 0,690 | 0,688 | 0,644 | 0,643 | 0,610 | 0,649 | 0,690 | 0,664 |
| Mlbqt.NV | | 1,065 | 1,134 | 1,056 | 1,058 | 1,063 | 0,956 | 1,018 | 1,086 | 1,033 |
| Mlbqt.Kat | | 1,068 | 1,131 | 1,044 | 1,015 | 1,081 | 0,981 | 1,022 | 1,084 | 1,013 |
| Linkage | | 0,399 | 0,014 | 0,170 | 1,040 | 0,229 | 0,823 | 1,049 | 0,485 | 0,920 |
| Gh.HE.Trad | Abhängig | 1,063 | 1,080 | 1,103 | 1,088 | 1,122 | 1,094 | 1,092 | 1,075 | 1,036 |
| Sage.HE | | 1,212 | 1,119 | 1,286 | 1,174 | 1,131 | 1,167 | 1,163 | 1,108 | 1,089 |
| rHE | | 1,662 | 1,531 | 1,452 | 1,356 | 1,342 | 1386 | 1,255 | 1,226 | 1,210 |
| Merlin-Regress | | 1,323 | 1,256 | 1,247 | 1,215 | 1,248 | 1,260 | 1,214 | 1,195 | 1,156 |
| Gh.VC | | 1,413 | 1,097 | 1,372 | 1,113 | 0,904 | 0,919 | 1,336 | 1,134 | 1,325 |
| Solar.VC | | 1,375 | 1,057 | 1,345 | 1,105 | 0,895 | 0,911 | 1,293 | 1,111 | 1,306 |
| Npar | | 1,142 | 1,195 | 1,144 | 1,065 | 1,196 | 1,156 | 1,131 | 1,135 | 1,134 |
| Merlin.K&C | | 0,954 | 1,022 | 0,938 | 0,930 | 0,995 | 0,979 | 0,930 | 0,999 | 0,908 |
| Merlin.W&H | | 0,721 | 0,765 | 0,681 | 0,669 | 0,729 | 0,734 | 0,699 | 0,725 | 0,667 |
| Mlbqt.NV | | 0,989 | 0,986 | 0,970 | 0,987 | 1,011 | 1,037 | 0,932 | 0,996 | 0,949 |
| Mlbqt.Kat | | 1,000 | 1,030 | 0,987 | 0,971 | 0,981 | 1,057 | 0,938 | 1,008 | 0,931 |
| Linkage | | 0,483 | 0,015 | 0,236 | 0,915 | 0,147 | 0,702 | 0,812 | 0,139 | 0,593 |

**Tabelle 26**: Empirischer Typ I-Fehler [%] bei einem nominalen Fehlerniveau von 0,1% unter Normalverteilungsannahmen.

| | | Empirischer Typ I-Fehler unter Normalverteilungsannahmen | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Gh.HE.Trad | Unabhängig | 0,087 | 0,103 | 0,101 | 0,101 | 0,094 | 0,090 | 0,117 | 0,097 | 0,110 |
| Sage.HE | | 0,083 | 0,098 | 0,097 | 0,099 | 0,089 | 0,088 | 0,115 | 0,092 | 0,104 |
| rHE | | 0,089 | 0,092 | 0,099 | 0,072 | 0,098 | 0,089 | 0,114 | 0,102 | 0,113 |
| Merlin-Regress | | 0,071 | 0,087 | 0,088 | 0,101 | 0,088 | 0,090 | 0,102 | 0,110 | 0,107 |
| Gh.VC | | 0,170 | 0,128 | 0,166 | 0,023 | 0,004 | 0,001 | 0,353 | 0,336 | 0,342 |
| Solar.VC | | 0,168 | 0,127 | 0,161 | 0,023 | 0,004 | 0,001 | 0,355 | 0,329 | 0,335 |
| Npar | | 0,088 | 0,107 | 0,106 | 0,103 | 0,089 | 0,099 | 0,113 | 0,096 | 0,119 |
| Merlin.K&C | | 0,095 | 0,099 | 0,094 | 0,102 | 0,108 | 0,080 | 0,088 | 0,102 | 0,106 |
| Merlin.W&H | | 0,056 | 0,059 | 0,050 | 0,059 | 0,060 | 0,039 | 0,043 | 0,062 | 0,057 |
| Mlbqt.NV | | 0,108 | 0,114 | 0,102 | 0,092 | 0,105 | 0,104 | 0,092 | 0,112 | 0,104 |
| Mlbqt.Kat | | 0,105 | 0,114 | 0,106 | 0,095 | 0,107 | 0,109 | 0,092 | 0,106 | 0,092 |
| Linkage | | 0,002 | 0,000 | 0,000 | 0,056 | 0,002 | 0,011 | 0,069 | 0,001 | 0,036 |
| Gh.HE.Trad | Abhängig | 0,116 | 0,135 | 0,124 | 0,126 | 0,121 | 0,150 | 0,123 | 0,119 | 0,109 |
| Sage.HE | | 0,170 | 0,154 | 0,179 | 0,150 | 0,139 | 0,167 | 0,148 | 0,144 | 0,120 |
| rHE | | 0,293 | 0,256 | 0,279 | 0,238 | 0,211 | 0,234 | 0,194 | 0,191 | 0,149 |
| Merlin-Regress | | 0,177 | 0,140 | 0,170 | 0,169 | 0,166 | 0,161 | 0,165 | 0,149 | 0,146 |
| Gh.VC | | 0,180 | 0,115 | 0,159 | 0,105 | 0,050 | 0,050 | 0,167 | 0,139 | 0,152 |
| Solar.VC | | 0,165 | 0,103 | 0,161 | 0,102 | 0,050 | 0,049 | 0,167 | 0,131 | 0,146 |
| Npar | | 0,141 | 0,113 | 0,138 | 0,115 | 0,127 | 0,144 | 0,129 | 0,134 | 0,131 |
| Merlin.K&C | | 0,103 | 0,101 | 0,106 | 0,079 | 0,100 | 0,101 | 0,099 | 0,115 | 0,089 |
| Merlin.W&H | | 0,065 | 0,058 | 0,067 | 0,048 | 0,061 | 0,069 | 0,067 | 0,066 | 0,059 |
| Mlbqt.NV | | 0,100 | 0,095 | 0,099 | 0,082 | 0,098 | 0,091 | 0,095 | 0,111 | 0,072 |
| Mlbqt.Kat | | 0,096 | 0,092 | 0,097 | 0,074 | 0,106 | 0,089 | 0,089 | 0,104 | 0,076 |
| Linkage | | 0,008 | 0,000 | 0,000 | 0,039 | 0,000 | 0,019 | 0,018 | 0,000 | 0,007 |

## 8.3.2 Unter Verletzung der Normalverteilungsannahmen

**Tabelle 27**: Empirischer Typ I-Fehler [%] bei einem nominalen Fehlerniveau von 5% unter Verletzung der Normalverteilungsannahmen.

| | | Empirischer Typ I-Fehler unter Verletzung der Normalverteilungsannahmen | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Gh.HE.Trad | Unabhängig | 5,00 | 5,16 | 5,07 | 5,01 | 5,07 | 4,99 | 5,19 | 5,13 | 5,07 |
| Sage.HE | | 4,94 | 5,09 | 5,01 | 4,95 | 5,00 | 4,90 | 5,11 | 5,07 | 5,01 |
| rHE | | 5,04 | 4,93 | 4,86 | 4,96 | 5,00 | 4,97 | 5,20 | 4,93 | 4,96 |
| Merlin-Regress | | 5,00 | 4,98 | 5,01 | 4,97 | 4,85 | 4,94 | 5,18 | 5,02 | 4,98 |
| Gh.VC | | 14,08 | 13,29 | 12,67 | 3,51 | 2,80 | 1,62 | 12,86 | 12,97 | 11,97 |
| Solar.VC | | 14,07 | 13,30 | 12,65 | 3,48 | 2,79 | 1,60 | 12,79 | 13,00 | 11,94 |
| Npar | | 4,93 | 4,98 | 5,02 | 5,13 | 4,91 | 5,13 | 4,96 | 5,09 | 5,00 |
| Merlin.K&C | | 4,47 | 4,49 | 4,63 | 4,59 | 4,60 | 4,45 | 4,61 | 4,59 | 4,54 |
| Merlin.W&H | | 3,60 | 3,63 | 3,68 | 3,65 | 3,73 | 3,56 | 3,69 | 3,72 | 3,62 |
| Mlbqt.NV | | 5,26 | 5,29 | 5,22 | 5,13 | 5,16 | 5,09 | 5,16 | 5,09 | 4,99 |
| Mlbqt.Kat | | 5,13 | 5,22 | 5,12 | 4,99 | 5,12 | 5,03 | 5,08 | 5,07 | 4,99 |
| Linkage | | 5,27 | 2,12 | 4,79 | 5,27 | 4,68 | 5,58 | 5,32 | 4,74 | 5,61 |
| Gh.HE.Trad | Abhängig | 5,17 | 5,15 | 5,27 | 5,29 | 5,21 | 5,33 | 5,21 | 5,17 | 5,20 |
| Sage.HE | | 4,48 | 4,16 | 4,68 | 3,55 | 3,36 | 3,68 | 4,56 | 4,24 | 4,67 |
| rHE | | 6,22 | 6,08 | 6,15 | 5,57 | 5,44 | 5,43 | 5,65 | 5,44 | 5,44 |
| Merlin-Regress | | 5,53 | 5,51 | 5,54 | 5,50 | 5,40 | 5,50 | 5,53 | 5,46 | 5,47 |
| Gh.VC | | 11,31 | 10,57 | 10,49 | 6,70 | 6,56 | 5,65 | 10,11 | 9,78 | 9,44 |
| Solar.VC | | 11,28 | 10,55 | 10,42 | 6,70 | 6,51 | 5,57 | 10,09 | 9,71 | 9,43 |
| Npar | | 5,27 | 5,25 | 5,19 | 5,30 | 5,16 | 5,25 | 5,25 | 5,19 | 5,17 |
| Merlin.K&C | | 4,49 | 4,50 | 4,49 | 4,52 | 4,52 | 4,49 | 4,57 | 4,56 | 4,63 |
| Merlin.W&H | | 3,76 | 3,76 | 3,75 | 3,75 | 3,78 | 3,70 | 3,85 | 3,81 | 3,92 |
| Mlbqt.NV | | 5,05 | 5,02 | 5,10 | 4,96 | 5,09 | 5,02 | 4,90 | 4,98 | 5,05 |
| Mlbqt.Kat | | 4,88 | 4,99 | 4,98 | 4,94 | 5,06 | 4,95 | 4,77 | 4,97 | 5,03 |
| Linkage | | 4,80 | 2,02 | 4,49 | 5,15 | 3,86 | 5,44 | 5,15 | 3,41 | 5,13 |

**Tabelle 28**: Empirischer Typ I-Fehler [%] bei einem nominalen Fehlerniveau von 1% unter Verletzung der Normalverteilungsannahmen.

| | | Empirischer Typ I-Fehler unter Verletzung der Normalverteilungsannahmen | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Gh.HE.Trad | Unabhängig | 0,620 | 0,646 | 0,620 | 0,669 | 0,654 | 0,683 | 0,643 | 0,648 | 0,622 |
| Sage.HE | | 0,672 | 0,716 | 0,688 | 0,735 | 0,721 | 0,709 | 0,728 | 0,712 | 0,667 |
| rHE | | 0,971 | 0,879 | 0,895 | 0,914 | 0,915 | 0,907 | 0,971 | 0,920 | 0,946 |
| Merlin-Regress | | 0,805 | 0,811 | 0,818 | 0,747 | 0,816 | 0,810 | 0,850 | 0,848 | 0,892 |
| Gh.VC | | 6,260 | 5,749 | 5,163 | 0,365 | 0,256 | 0,100 | 5,486 | 5,685 | 4,880 |
| Solar.VC | | 6,270 | 5,779 | 5,193 | 0,364 | 0,247 | 0,097 | 5,481 | 5,645 | 4,884 |
| Npar | | 0,989 | 0,961 | 0,929 | 0,972 | 0,954 | 1,075 | 0,978 | 0,969 | 0,967 |
| Merlin.K&C | | 0,957 | 1,019 | 0,945 | 0,961 | 0,970 | 0,966 | 1,009 | 0,950 | 1,007 |
| Merlin.W&H | | 0,666 | 0,704 | 0,634 | 0,665 | 0,649 | 0,673 | 0,711 | 0,629 | 0,687 |
| Mlbqt.NV | | 1,044 | 1,053 | 1,082 | 1,077 | 1,008 | 0,986 | 1,037 | 0,967 | 1,055 |
| Mlbqt.Kat | | 1,043 | 1,025 | 1,038 | 1,043 | 0,975 | 1,009 | 1,039 | 0,964 | 1,014 |
| Linkage | | 0,580 | 0,021 | 0,288 | 1,062 | 0,304 | 1,059 | 1,081 | 0,368 | 0,932 |
| Gh.HE.Trad | Abhängig | 0,885 | 0,865 | 0,957 | 0,983 | 0,939 | 0,975 | 0,886 | 0,909 | 0,929 |
| Sage.HE | | 1,108 | 0,955 | 1,137 | 0,732 | 0,682 | 0,781 | 1,169 | 0,975 | 1,104 |
| rHE | | 1,860 | 1,700 | 1,724 | 1,410 | 1,319 | 1,327 | 1,451 | 1,328 | 1,351 |
| Merlin-Regress | | 1,293 | 1,230 | 1,250 | 1,185 | 1,193 | 1,214 | 1,212 | 1,197 | 1,200 |
| Gh.VC | | 4,764 | 4,130 | 3,979 | 1,525 | 1,432 | 1,060 | 3,947 | 3,748 | 3,427 |
| Solar.VC | | 4,739 | 4,096 | 3,971 | 1,514 | 1,451 | 1,070 | 3,951 | 3,745 | 3,391 |
| Npar | | 1,106 | 1,179 | 1,116 | 1,208 | 1,142 | 1,218 | 1,132 | 1,080 | 1,118 |
| Merlin.K&C | | 0,941 | 0,971 | 0,946 | 0,938 | 0,927 | 0,945 | 0,946 | 0,997 | 1,011 |
| Merlin.W&H | | 0,734 | 0,712 | 0,724 | 0,695 | 0,673 | 0,697 | 0,720 | 0,744 | 0,750 |
| Mlbqt.NV | | 0,983 | 0,966 | 1,049 | 0,985 | 0,996 | 0,985 | 1,024 | 0,978 | 1,007 |
| Mlbqt.Kat | | 0,971 | 0,969 | 1,019 | 0,961 | 1,021 | 0,972 | 0,984 | 1,003 | 0,997 |
| Linkage | | 0,593 | 0,031 | 0,360 | 0,984 | 0,213 | 0,861 | 0,882 | 0,159 | 0,659 |

**Tabelle 29**: Empirischer Typ I-Fehler [%] bei einem nominalen Fehlerniveau von 0,1% unter Verletzung der Normalverteilungsannahmen.

| | | Empirischer Typ I-Fehler unter Verletzung der Normalverteilungsannahmen | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Geschwister- | Ohne Selektion | | | Einfache Selektion | | | Doppelte Selektion | | |
| Verfahren | schaften | Dom | Add | Rez | Dom | Add | Rez | Dom | Add | Rez |
| Gh.HE.Trad | Unabhängig | 0,030 | 0,028 | 0,050 | 0,029 | 0,043 | 0,041 | 0,040 | 0,047 | 0,031 |
| Sage.HE | | 0,028 | 0,028 | 0,049 | 0,029 | 0,040 | 0,035 | 0,037 | 0,045 | 0,028 |
| rHE | | 0,096 | 0,077 | 0,093 | 0,088 | 0,074 | 0,072 | 0,076 | 0,088 | 0,078 |
| Merlin-Regress | | 0,061 | 0,053 | 0,058 | 0,067 | 0,049 | 0,059 | 0,058 | 0,074 | 0,064 |
| Gh.VC | | 2,009 | 1,801 | 1,330 | 0,019 | 0,019 | 0,012 | 1,821 | 2,014 | 1,399 |
| Solar.VC | | 2,007 | 1,799 | 1,328 | 0,018 | 0,018 | 0,012 | 1,817 | 2,008 | 1,389 |
| Npar | | 0,107 | 0,094 | 0,104 | 0,081 | 0,099 | 0,089 | 0,081 | 0,088 | 0,103 |
| Merlin.K&C | | 0,098 | 0,102 | 0,113 | 0,113 | 0,084 | 0,089 | 0,097 | 0,079 | 0,106 |
| Merlin.W&H | | 0,050 | 0,053 | 0,067 | 0,067 | 0,046 | 0,053 | 0,058 | 0,038 | 0,062 |
| Mlbqt.NV | | 0,103 | 0,105 | 0,108 | 0,100 | 0,100 | 0,113 | 0,109 | 0,069 | 0,121 |
| Mlbqt.Kat | | 0,109 | 0,087 | 0,101 | 0,094 | 0,091 | 0,105 | 0,105 | 0,067 | 0,120 |
| Linkage | | 0,003 | 0,000 | 0,001 | 0,058 | 0,002 | 0,045 | 0,055 | 0,001 | 0,029 |
| Gh.HE.Trad | Abhängig | 0,073 | 0,059 | 0,094 | 0,081 | 0,094 | 0,098 | 0,090 | 0,096 | 0,089 |
| Sage.HE | | 0,199 | 0,163 | 0,195 | 0,087 | 0,106 | 0,102 | 0,191 | 0,162 | 0,182 |
| rHE | | 0,376 | 0,380 | 0,370 | 0,238 | 0,189 | 0,214 | 0,272 | 0,237 | 0,209 |
| Merlin-Regress | | 0,143 | 0,166 | 0,169 | 0,144 | 0,126 | 0,142 | 0,158 | 0,128 | 0,165 |
| Gh.VC | | 1,430 | 1,187 | 0,991 | 0,171 | 0,144 | 0,104 | 1,254 | 1,242 | 0,946 |
| Solar.VC | | 1,434 | 1,191 | 0,981 | 0,166 | 0,141 | 0,105 | 1,239 | 1,237 | 0,927 |
| Npar | | 0,125 | 0,134 | 0,136 | 0,150 | 0,141 | 0,149 | 0,128 | 0,126 | 0,130 |
| Merlin.K&C | | 0,104 | 0,087 | 0,091 | 0,091 | 0,086 | 0,098 | 0,098 | 0,088 | 0,098 |
| Merlin.W&H | | 0,063 | 0,060 | 0,061 | 0,055 | 0,057 | 0,070 | 0,070 | 0,060 | 0,054 |
| Mlbqt.NV | | 0,084 | 0,088 | 0,085 | 0,098 | 0,103 | 0,083 | 0,112 | 0,105 | 0,102 |
| Mlbqt.Kat | | 0,075 | 0,099 | 0,093 | 0,088 | 0,104 | 0,083 | 0,097 | 0,106 | 0,092 |
| Linkage | | 0,006 | 0,000 | 0,002 | 0,043 | 0,000 | 0,017 | 0,033 | 0,000 | 0,014 |

## 8.4 Empirische Typ I-Fehler und Power bei Missspezifikation der Modelparameter für Merlin-Regress

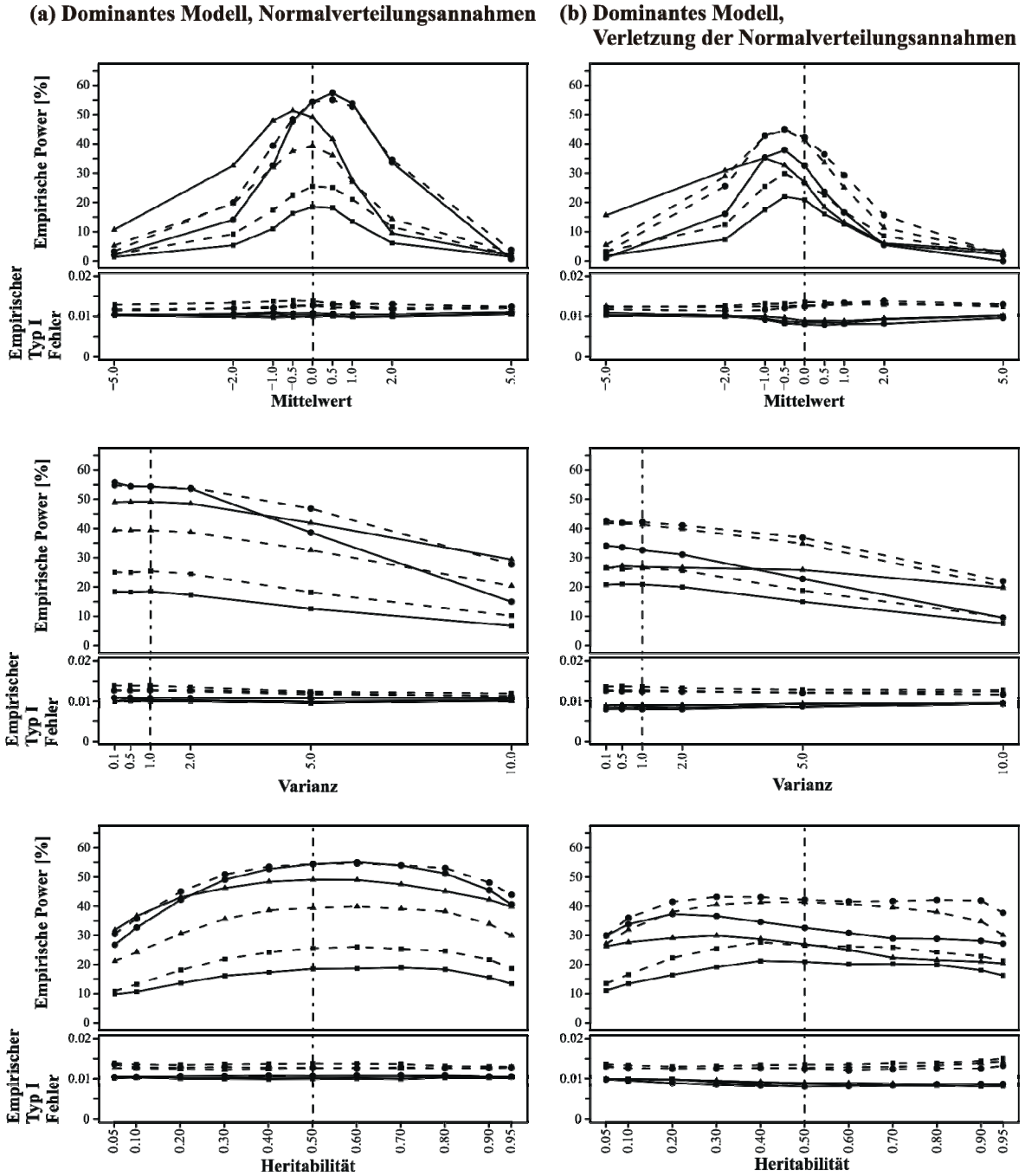


**Abbildung 6**: Effekt der Model-Missspezifikation auf den empirischen Typ I-Fehler und Power für ein dominantes genetisches Modell unter (a) Normalverteilungsannahmen und (b) Verletzung der Normalverteilungsannahmen für das MERLIN-REGRESS Verfahren. Der empirische Typ I-Fehler wurde bei einem nominalen Typ I-Fehler von 0,01 errechnet, während die Power bei einem empirischen Typ I–Fehler von 0,01 berechnet worden ist. Die durchgezeichneten Linien zeigen unabhängige Geschwisterschaften, während die gestrichelten Linien abhängige Geschwisterschaften zeigen. Drei Selektionsschema werden gezeigt: ■ zufällige Selektion, ● einfache Selektion, ▲ doppelte Selektion. Das wahre populationsbasierte Modell ist mit den senkrechten gestrichelten Linien eingezeichnet.

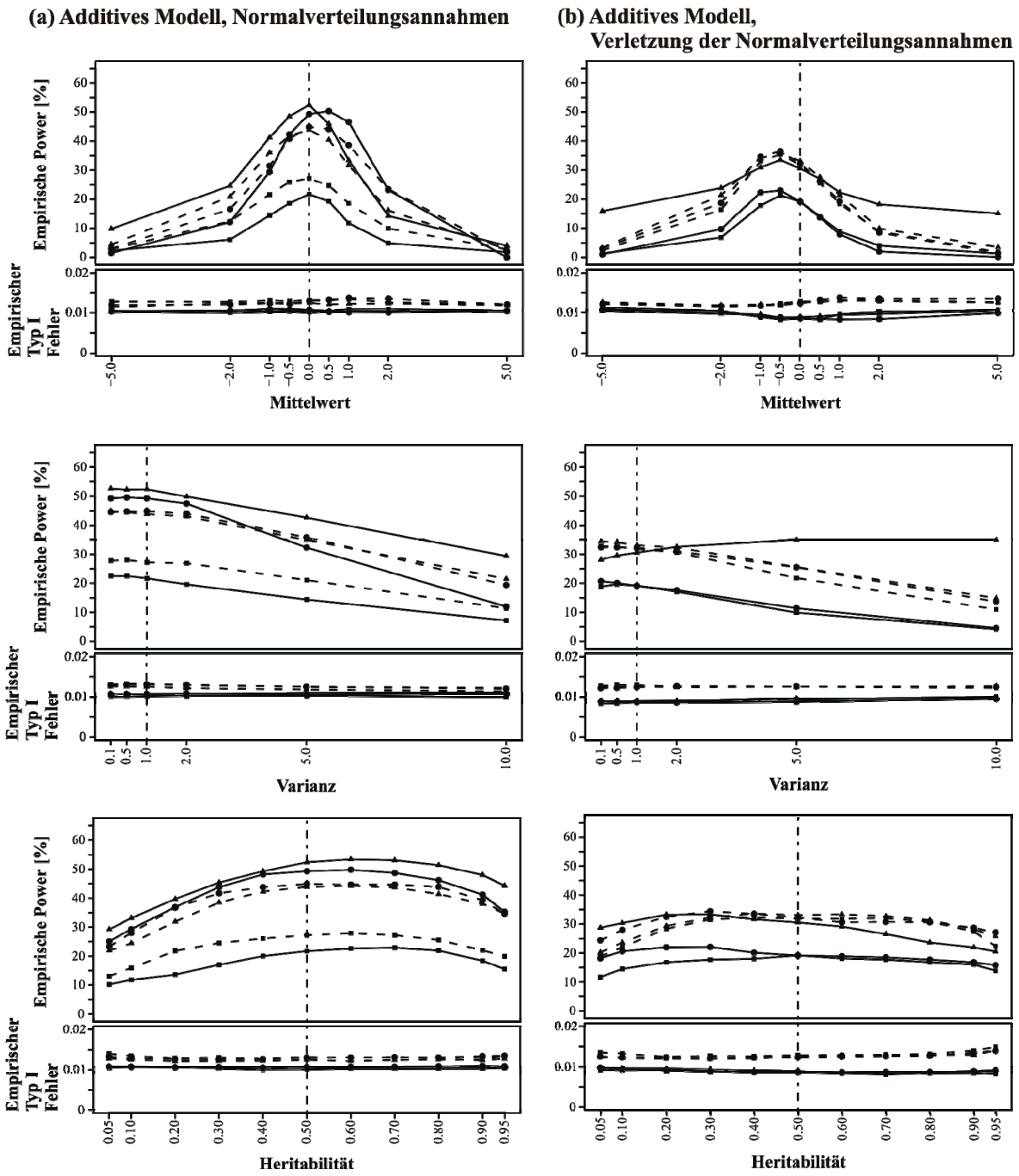


**Abbildung 7**: Effekt der Model-Missspezifikation auf den empirischen Typ I-Fehler und Power für ein additives genetisches Modell unter (a) Normalverteilungsannahmen und (b) Verletzung der Normalverteilungsannahmen für das MERLIN-REGRESS Verfahren. Der empirische Typ I-Fehler wurde bei einem nominalen Typ I-Fehler von 0,01 errechnet, während die Power bei einem empirischen Typ I–Fehler von 0,01 berechnet worden ist. Die durchgezeichneten Linien zeigen unabhängige Geschwisterschaften, während die gestrichelten Linien abhängige Geschwisterschaften zeigen. Drei Selektionsschema werden gezeigt: ■ zufällige Selektion, ● einfache Selektion, ▲ doppelte Selektion. Das wahre populationsbasierte Modell ist mit den senkrechten gestrichelten Linien eingezeichnet.

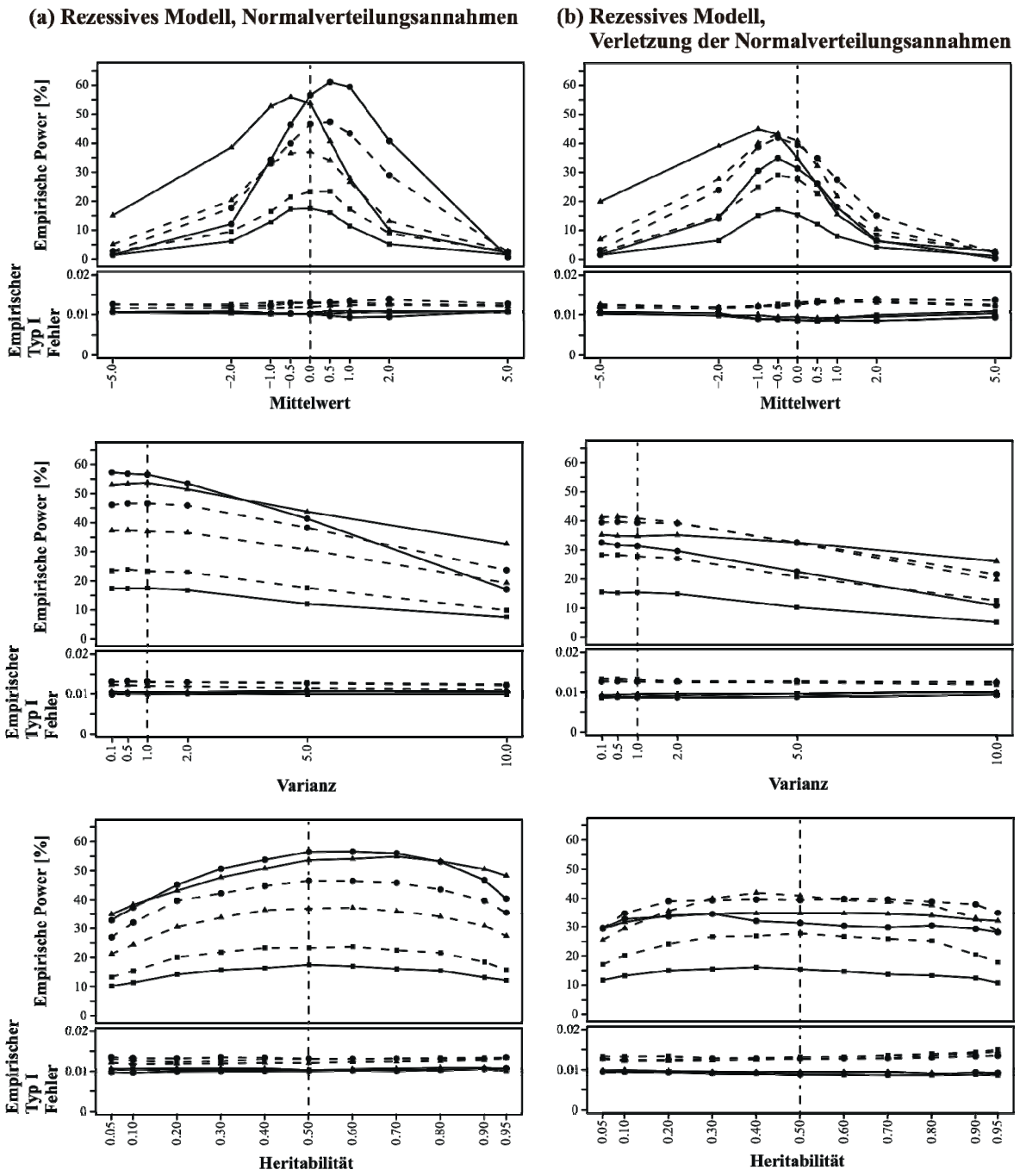


**Abbildung 8**: Effekt der Model-Missspezifikation auf den empirischen Typ I-Fehler und Power für ein rezessives genetisches Modell unter (a) Normalverteilungsannahmen und (b) Verletzung der Normalverteilungsannahmen für das MERLIN-REGRESS Verfahren. Der empirische Typ I-Fehler wurde bei einem nominalen Typ I-Fehler von 0,01 errechnet, während die Power bei einem empirischen Typ I–Fehler von 0,01 berechnet worden ist. Die durchgezeichneten Linien zeigen unabhängige Geschwisterschaften, während die gestrichelten Linien abhängige Geschwisterschaften zeigen. Drei Selektionsschema werden gezeigt: ■ zufällige Selektion, ● einfache Selektion, ▲ doppelte Selektion. Das wahre populationsbasierte Modell ist mit den senkrechten gestrichelten Linien eingezeichnet.

# 9 Danksagungen

An erster Stelle gilt mein herzlicher Dank Herrn Prof. Dr. rer. nat. Ziegler für die engagierte Förderung meines Promotionsvorhabens und die Betreuung meiner wissenschaftlichen Tätigkeit am IMBS. Ihm verdanke ich wertvolle Anregungen, ohne die diese Arbeit nicht entstanden wäre.

Stellvertretend für alle Kollegen am IMBS danke ich Frau Dr. rer. hum. biol. König, die mir wie zahlreiche andere Mitarbeiter des Instituts stets mit Rat und Tat zur Verfügung stand. Das kollegiale Klima am IMBS und die mit meinen dortigen Kollegen geführten Gespräche zu fachlichen Themen haben maßgeblich zum Gelingen dieser Arbeit beigetragen.

Herrn Dr. Palmer danke ich für den in dieser Arbeit erneut analysierten COAG Perth Datensatz. Mein besonderer Dank gilt darüber hinaus Herrn Dr. rer. hum. biol. Franke, der mit mir das Softwarepaket `Sibsim` erstellt hat.

Schließlich möchte ich auch Herrn Dr. jur. Schlichte danken, der die mühevolle Arbeit des Korrekturlesens auf sich genommen hat.

# 10 Lebenslauf

| | |
|---|---|
| Name | André Kleensang |
| Anschrift | Pastorenstr. 7, 20459 Hamburg |
| Geburtsdatum/-ort | 24.01.1974 in Hamburg |
| Familienstand | ledig |
| Staatsangehörigkeit | deutsch |

| | |
|---|---|
| 1985-1990 | Schulausbildung an der Haupt- und Realschule Richard-Linde-Weg in Hamburg |
| 1990 | Realschulabschluss |
| 1990-1994 | Schulausbildung an der Staatlichen Gewerbeschule Chemie, Pharmazie, Agrarwirtschaft und Gesamtschule Bergedorf in Hamburg |
| 1994 | Allgemeine Hochschulreife |
| 1994 | Staatlich geprüfter chemisch-technischer Assistent |
| 1994-1995 | Grundwehrdienst als ABC-Aufklärungssoldat/Stabsdienstsoldat ABC-Abwehrbataillion 610 in Albersdorf |
| 1995-2001 | Studium der Biochemie an der Universität Hamburg |
| 1998 | Stipendium des Erasmus/Sokrates-Programms zum Studium der Biochemie an der Universität Bern, Schweiz |
| 2001 | Abschluss des Biochemiestudiums zum Diplom-Biochemiker |
| 2001 | Wissenschaftlicher Mitarbeiter am Bernhard-Nocht-Institut für Tropenmedizin in Hamburg, Abteilung für Molekulare Parasitologie (bei Dr. med. Klaus Erttmann) |
| 2002 | Wissenschaftlicher Mitarbeiter am Bernhard-Nocht-Institut für Tropenmedizin in Hamburg, Bioinformatics Research Lab (bei Dr. med. Bertram Müller-Myhsok) |
| 2003-2007 | Wissenschaftlicher Mitarbeiter am Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein / Campus Lübeck (bei Prof. Dr. rer. nat. Andreas Ziegler) |
| Seit 2003 | Promotionsstudium der Humanbiologie an der Universität zu Lübeck |
| Seit 2004 | Postgraduierten-Studium der Wirtschaftswissenschaften zum Diplom-Wirtschaftschemiker an der Fernuniversität Hagen (Vordiplom 2006) |
| Seit 2007 | Vertragsbediensteter bei der Europäischen Kommission, Gemeinsame Forschungsstelle, Institut für Gesundheit und Verbraucherschutz, Ispra, Italien |

# 11 Publikationsliste (Stand 01.02.2010)

Die aus dieser Dissertation resultierenden Publikationen sind mit einem Sternchen gekennzeichnet.

* **Kleensang A**, Franke D, Alcaïs A, Abel L, Müller-Myhsok B, Ziegler A (2010): An Extensive Comparison of Quantitative Trait Loci Mapping Methods. Hum Hered (im Druck)

Ziegler A, Ewhida A, Brendel M, **Kleensang A** (2008): More Powerful Haplotype Sharing by Accounting for the Mode of Inheritance. Genet Epidemiol 33(3):228-36

Timmann C, van der Kamp E, **Kleensang A**, König I K, Thye T, Büttner D W, Hamelmann C, Marfo Y, Vens M, Brattig N, Ziegler A, Horstmann R D (2008): Human Genetic Resistance to Onchocerca volvulus: Evidence for Linkage to Chromosome 2p from an Autosome-wide Scan. J Infect Dis 198(3):427-33

Lohmann-Hedrich K, Neumann A, **Kleensang A**, Lohnau T, Muhle H, Djarmati A, König IR, Pramstaller PP, Schwinger E, Kramer PL, Ziegler A, Stephani U, Klein C (2008): Evidence for linkage of restless legs syndrome to chromosome 9p: Are there two distinct loci? Neurology 70(9):686-94

**Kleensang A**, Pahlke F, Ziegler A (2007): Familienstudien in der Genetischen Epidemiologie: Ein Überblick. In Freyer G, Biebler K E (eds.): Biometrische Aspekte der Genomanalyse III, Shaker Verlag, Aachen, Germany, 3-20

Timmann C, Evans JA, König IR, **Kleensang A**, Rüschendorf F, Lenzen J, Sievertsen J, Becker C, Enuameh Y, Kwakye KO, Opoku E, Browne ENL, Ziegler A, Nürnberg P, Horstmann RD (2007): Genome-Wide Linkage Analysis of Malaria Infection Intensity and Mild Malaria Disease. PLoS Genet 3(3):e48

Schulte-Körne G, Ziegler A, Deimel W, Schumacher J, Plume E, Bachmann C, **Kleensang A**, Propping P, Nöthen MM, Warnke A, Remschmidt H, König IR (2007): Interrelationship and familiality of dyslexia related quantitative measures. Ann Hum Genet 71(Pt 2):160-75

* Franke D, **Kleensang A**, Ziegler A (2006): SIBSIM - quantitative phenotype simulation in extended pedigrees. GMS Med Inform Biom Epidemiol 2(1):Doc04

Schumacher J, Konig IR, Plume E, Propping P, Warnke A, Manthey M, Duell M, **Kleensang A**, Repsilber D, Preis M, Remschmidt H, Ziegler A, Nothen MM, Schulte-Körne G (2006):

Linkage analyses of chromosomal region 18q11-q12 in dyslexia. J Neural Transm 113(3):417-23

**Kleensang A**, Franke D, König IR, Ziegler A (2005): Haplotype sharing analysis for alcohol dependence based on quantitative traits and the Mantel statistic. BMC Genetics 6(Suppl 1):S75

Franke D, **Kleensang A**, Elston RC, Ziegler AZ (2005): Haseman-Elston weighted by marker informativity. BMC Genetics 6(Suppl 1):S50

Erttmann KD, **Kleensang A**, Schneider E, Hammerschmidt S, Büttner DW, Gallin M (2005): Cloning, characterization and DNA immunization of an Onchocerca volvulus glyceraldehyde-3-phosphate dehydrogenase (Ov-GAPDH). Biochim Biophys Acta 1741:85-94

Mossner R, Kingo K, **Kleensang A**, Krüger U, König IR, Silm H, Westphal GA, Reich K (2005): Association of TNF -238 and -308 Promoter Polymorphisms with Psoriasis Vulgaris and Psoriatic Arthritis but not with Pustulosis Palmoplantaris. J Invest Dermatol 124: 282-284

Ziegler A, König IR, Deimel W, Plume E, Nöthen MM, Propping P, **Kleensang A**, Müller-Myhsok B, Warnke A, Remschmidt H, Schulte-Körne G (2005): Developmental Dyslexia-recurrence risk estimates from a german bi-center study using the single proband sib pair design. Hum Hered 59:136-143

**Kleensang A**, König IR (2004): A novel implementation of a robust variance component approach exemplified using SOLAR. Genet Epidemiol 27:280

Hennies CH, **Kleensang A**, Blech H, Meyer B, Schmidt S, Ziegler A, McElwee K, Hoffmann R (2004): Genetic mapping in alopecia areata. J Deut Dermatol Ges 2:496

König IR, Ziegler A, Schumacher J, Nöthen MM, Plume E, **Kleensang A**, Warnke A, Propping P, Remschmidt H, Schulte-Körne G (2004): Linkage analyses on chromosomal regions 15q21 and 18p11 in dyslexia — results from the German bi-center study. Genet Epidemiol 27:281

König IR, Repsilber D, Dahmen G, **Kleensang A**, Ziegler A (2004): Anwendungsorientiertere Ausbildung im Teil "Medizinische Biometrie" des Querschnittsfachs Q1 durch Einbettung von Konzepten der Evidenzbasierten Medizin - Ein Erfahrungsbericht nach Umstellung auf die neue ÄAppO. Inform Biom Epidemiol Med Biol 35(4):220-228