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

Empfehlungen zur Qualitätssicherung von Genotypisierungsdaten bei familienbasierten Studien mit Mikrosatelliten

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck
- aus der medizinischen Fakultät -

vorgelegt von
Alexander Kiewert
aus Osnabrück

Lübeck 2006

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

2. Berichterstatter: Prof. Dr. rer. nat. Georg Sczakiel

Tag der mündlichen Prüfung: 17.07.2007

Zum Druck genehmigt. Lübeck, den 17.07.2007

Gez. Prof. Dr. med. Werner Solbach
- Dekan der Medizinischen Fakultät -

Meinem Vater gewidmet

Inhaltsverzeichnis

1 Einleitung

1.1 Ziel und Fragestellung der Arbeit

Im Jahre 1896 wurde von britischen Augenärzten von einem Fall von „kongenitaler Wortblindheit“ berichtet (Pringle-Morgan, 1896). Das beschriebene Phänomen ist heute als Entwicklungsstörung unter den Namen Dyslexie, Legasthenie oder Lese-Rechtschreibschwäche (LRS) bekannt.
Der Dyslexie liegt eine multifaktorielle Genese zugrunde, dass heißt, es gibt mehr als eine pathogenetische Ursache für die Krankheit. Mehrere Faktoren scheinen bei der Entstehung der Dyslexie eine Rolle zu spielen. Einer dieser Faktoren ist eine erbliche Komponente (Remschmidt et al., 2000).

Bisher sind ca. 5.000 Erbkrankheiten bekannt, die durch Veränderungen in einem einzigen Gen verursacht werden (OMIM). Darüber hinaus spielt die genetische Prädisposition eine wichtige Rolle für häufige Erkrankungen wie z. B. Bluthochdruck, Fettstoffwechselstörungen, Diabetes, Allergien und Tumorerkrankungen.
Bei Erkrankungen, wie bei der Dyslexie wird intensiv nach dem möglichen Locus gesucht, der die genetische Komponente der Dyslexie darstellt.

Diese Arbeit ist im Rahmen einer bizentrischen Studie zur Lese- und Rechtschreibschwäche (LRS-Studie) entstanden. Grundlage ist die Phänotypisierung von betroffenen Probanden und ihrer Geschwister sowie die Genotypisierung (Genomscan) der Kinder und ihrer Eltern.
Ziel der Studie ist es, bisherige Studienergebnisse zur molekulargenetischen Ursache der Dyslexie zu bestätigen und eventuelle neue Loci zu entdecken.
Die Suche nach einem möglichen Locus für Dyslexie wurde mittels einer genomweiten Kopplungsanalyse durchgeführt. Kopplungsanalysen stellen eine Möglichkeit dar, unbekannte Loci eines krankheitsauslösenden Gens zu bestimmen. Eine weitere Möglichkeit stellen Assoziationsstudien dar. Die Prinzipien der beiden Ansätze sind zum Beispiel bei Ott oder bei Strachan und Read erklärt (Ott, 1999; Strachan & Read, 2004).

In jeder Untersuchung, die mit Genotypisierungsdaten arbeitet, treten Fehler in den Daten auf. Diese Fehler sind nicht einfach zu vermeiden oder zu finden (Sobel et al., 2002).
Ein Beispiel dafür, dass solche Fehler auftreten und oft weitreichende Folgen haben können, ist der *Nature* Artikel von Gagneux et al. aus dem Jahr 1997 (Gagneux, 1997). Die

Forschergruppe um Gagneux arbeitete mit Genotypisierungsdaten von Schimpansen. Aufgrund der Studienergebnisse war sie davon ausgegangen, dass sich weibliche Schimpansen regelmäßig außerhalb ihrer sozialen Gruppe „heimlich“ sexuell betätigten. Mehr als die Hälfte der Schimpansenkinder hätten Väter haben müssen, die nicht aus derselben Gruppe wie die Mütter stammten.

Dieser Bericht wurde von der allgemeinen Presse natürlich gerne aufgegriffen (Angier, 1997). Vier Jahre später musste Gagneux seine Ergebnisse zurückziehen. Sie beruhten auf falschen Schlussfolgerungen infolge unentdeckter Genotypisierungsfehler (Angier, 2001).

Dieses Beispiel zeigt, dass vor der Analyse der erhobenen Daten stets eine Kontrolle auf mögliche Fehler in den Datensätzen stattfinden sollte. Selbst wenn die Daten auf Fehler untersucht werden, wird man dabei nicht alle entdecken können.

In dieser Arbeit sollen Methoden beschrieben werden, die zum Entdecken von Fehlern in Genotypisierungsdaten geeignet sind. Diese Methoden sollen auf die LRS-Studie angewandt werden.

So kann diese Arbeit als „Handbuch zur Qualitätssicherung“ für Genotypisierungsdaten mit Mikrosatelliten in Familienstudien nach ihrer Erfassung im Labor und vor der weiteren Analyse in Kopplungsuntersuchungen verstanden werden.

1.2 Aufbau der Arbeit

Im Kapitel 2 wird näher auf die Klinik der Dyslexie eingegangen, und es werden die Ansätze zum Studiendesign und zur Phänotypisierung der untersuchten Personen erläutert. Ein Überblick über bisherige Studien zur Genetik der Dyslexie schließt dieses Kapitel ab.

Im dritten Kapitel werden die Hintergründe der LRS-Studie behandelt. In dieser Arbeit werden die Genotypisierungsdaten der LRS-Studie als „Material“ benutzt. Um den inhaltlichen Fluss nicht zu sehr zu stören, steht dieses Kapitel direkt hinter dem Kapitel Dyslexie.

Im Kapitel 4 geht es um die Fehler in molekulargenetischen Untersuchungen, und es wird erklärt, was genetische Marker – die Träger dieser Fehler – sind. Es werden die beiden grundsätzlichen Arten von Fehlern und Ansätze, diese Fehler zu entdecken, vorgestellt.

Kapitel 5, 6, 7, 8 und 9 behandeln die Methoden, mit denen die Genotypisierungsdaten auf Fehler untersucht werden. Zunächst werden im Kapitel 5 die wichtigsten Dateiformate vorgestellt, mit denen die Daten aus der LRS-Studie ausgewertet werden. Im Kapitel 6 ist eine

Methode beschrieben, die die Familienstruktur nutzt, um Fehler in den Daten zu entdecken. Im Kapitel 7 wird eine Möglichkeit vorgestellt, mit der die Daten auf Doppelrekombinationen als Ausdruck von Genotypisierungsfehlern untersucht werden können. Wie bei der Kartierung Fehler entdeckt und vermieden werden können, ist im Kapitel 8 beschrieben. Im Kapitel 9 werden verschiedenen Methoden vorgestellt, mit denen die Daten auf Abweichungen von der Hardy-Weinberg Verteilung untersucht werden können, und es wird erläutert, wie diese Abweichungen auf Fehler deuten können.

Eine Zusammenfassung der Methoden zur Qualitätssicherung stellt das Kapitel 10 dar. Dieses Kapitel ist in Form einer Standardarbeitsanweisung abgefasst. Diese Standardarbeits-anweisung ist so verfasst, das sie eine in sich geschlossene Darstellung der Vorgehensweise einer Qualitätssicherung bei Studien mit Genotypisierungsdaten enthält. Für zukünftige Studien kann das Kapitel 10 direkt als fertige Standardarbeitsanweisung zur Qualitätssicherung dienen.

Im Kapitel 11 werden die Ergebnisse dargelegt, die die vorgestellten Methoden angewandt auf die Genotypisierungsdaten der LRS-Studie ergaben.

Diese Arbeit schließt mit einer Diskussion der Ergebnisse der Qualitätssicherung und der Problematik der Entdeckung von Fehlern in Daten bei molekulargenetischen Untersuchungen allgemein in Kapitel 12. Hier wird auch diskutiert, wann welche Methoden sinnvoll sind und angewandt werden sollen.

2 Dyslexie

2.1 Klinische Aspekte der Dyslexie

Die Dyslexie, auch Legasthenie oder Lese-Rechtschreibschwäche (LRS), wurde erstmals im 19. Jahrhundert beobachtet. Britische Augenärzte beschrieben sie als kongenitale Wortblindheit (Pringle-Morgan, 1896). Vorher war die Dyslexie wohl unentdeckt geblieben, weil erst mit Einführung der Schulpflicht Lesen und Schreiben für die gesamte Bevölkerung wichtig wurden. Die Dyslexie ist nicht schicht- oder ortsgebunden und tritt in allen Schriftsprachen auf. Der Begriff Dyslexie stammt aus dem Griechischen und bedeutet übersetzt so viel wie „schlechte Sprache“ (dys: schlecht, lexis: Sprache). Die im deutschen synonym benutzte Legasthenie ist die „Schwäche beim Lesen“ (legein: lesen, astheneia: Schwäche).

Heute zählt die LRS zu den häufigsten kinder- und jugendpsychiatrischen Störungen. Das häufige Auftreten lässt sich dadurch erklären, dass es bis zur Einführung der Schulpflicht praktisch keinen Selektionsdruck gab. Die Dyslexie weist eine Prävalenz von 4% bis 9% auf (Shaywitz & Shaywitz et al., 1990; Schulte-Körne, 2001a; Francks, MacPhie et al., 2002), wobei Jungen im Durchschnitt 2-3-mal häufiger betroffen sind als Mädchen (Katusic et al., 2001; Rutter et al., 2004). Die Erstdiagnose wird in Deutschland meist in der zweiten Klasse gestellt. Dyslexie beeinträchtigt die Betroffenen nachhaltig. So ergaben Untersuchungen, dass Betroffene im Durchschnitt einen schlechteren Schulabschluss machen. Sie erreichen ein geringeres Berufsausbildungsniveau und sind im Vergleich zu Nicht-Betroffenen häufiger arbeitslos (Strehlow et al., 1992; Esser et al., 2002).

Nach der Einteilung der Weltgesundheitsorganisation, ICD-10, wird die Dyslexie als eine umschriebene Entwicklungsstörung bezeichnet, die nicht durch das Entwicklungsalter, Visusprobleme oder unangemessene Beschulung erklärbar ist (WHO, 1992). Neben der Dyslexie als kombinierte Lese-Rechtschreibschwäche (F81.0) ist auch eine Form der alleinigen Rechtschreibschwäche anerkannt (F81.1). Klinisch ist auch eine alleinige Leseschwäche beschrieben worden.

Das klinische Bild der Dyslexie zeigt sich beim Lesen und Schreiben. So haben betroffene Kinder Schwierigkeiten, Buchstaben, Wortteile und Worte korrekt zu lesen und Gelesenes wieder zu erkennen. Wortteile oder Worte werden oft ausgelassen, hinzugefügt, verdreht oder vertauscht. Textstellen gehen verloren und können nicht wiedergefunden werden. Der Textzusammenhang wird nicht verstanden und es können keine Schlussfolgerungen aus zuvor Gelesenem gezogen werden. Das Lesetempo ist deutlich verlangsamt. Bei der

Rechtschreibung zeigen sich Schwierigkeiten beim mündlichen Buchstabieren und beim Schreiben von Buchstaben, Wörtern und Sätzen. Bei Diktaten und beim Abschreiben von Texten werden viele Fehler gemacht (Schulte-Körne & Remschmidt, 2003). Als Folge der Dyslexie kann es zu begleitenden emotionalen Störungen und Verhaltensauffälligkeiten wie Herumalbern, Aggression, Traurigkeit und Angst vor der Schule kommen (Warnke & Roth, 2000).

LRS ist eine komplexe Erkrankung. Das heißt, dass nicht nur eine Ursache zugrunde liegt. Bisherige Untersuchungen weisen unter anderem auch auf eine genetische Komponente hin. Weiter sind kognitive Funktionen und eventuell auch Umweltfaktoren verantwortlich. Abbildung 2.1 gibt einen Überblick über mögliche ursächliche Komponenten der Dyslexie.

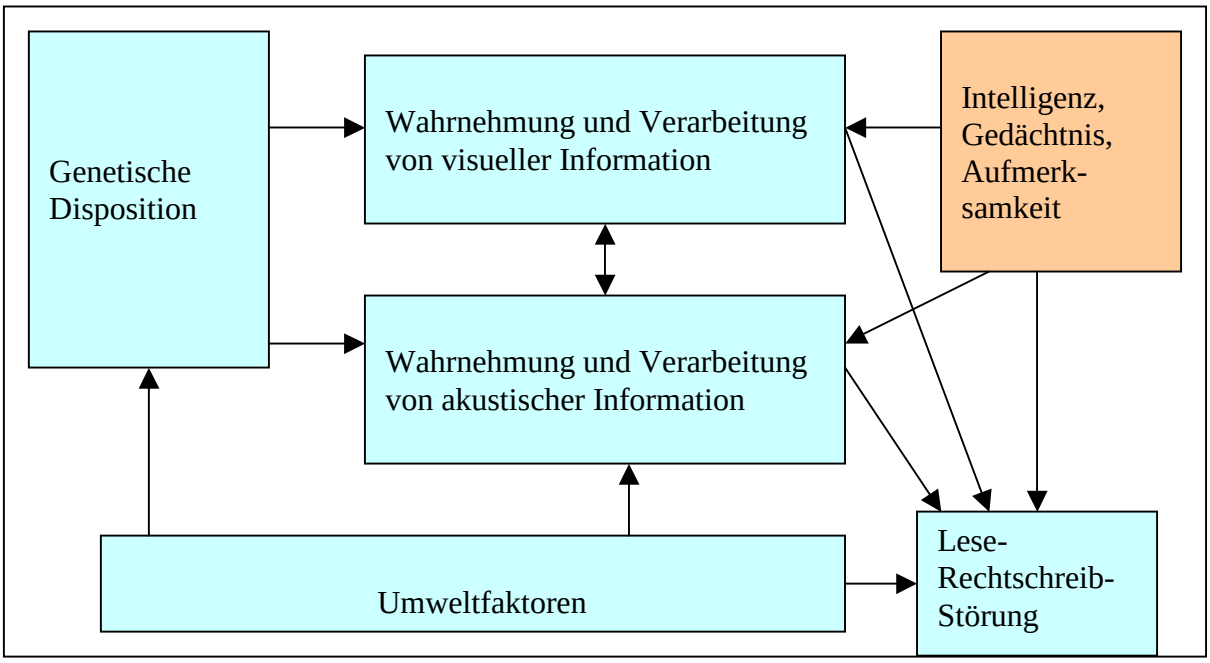

Abbildung 2.1
Mehrebenen-Ursachen-Modell – nach Schulte-Körne und Remschmidt (2003)

Für eine genetische Disposition zur LRS sprechen viele bisherige Untersuchungen. So zeigte sich in mehreren Studien eine familiäre Häufung. Das Wiederholungsrisiko für Geschwister liegt zwischen 38–62% (Schulte-Körne, 1998a; Schulte-Körne, 2001b; Ziegler et al., 2005). Eine neuere Studie aus dem arabischen Raum zeigte, dass Dyslexie häufiger unter Kindern nah blutsverwandter Eltern (*First-cousin parents*) als unter Kinder weiter entfernt verwandter Eltern auftritt, was auf eine genetische Komponente der Dyslexie deutet (Abu-Rabia & Maroun, 2005). Auf eine solche weisen ebenso Zwillingsuntersuchungen. Für die

Lesefähigkeit liegt der genetische Anteil an der beobachteten Gesamtvarianz zwischen 3% und 60%, für die Rechtschreibung bei 60-70 % (Stevenson, 1991; Olson et al., 1994). Bei Jungen scheint der genetische Einfluss höher zu sein als bei Mädchen.
Da die genetische Komponente der Dyslexie eindeutig ist, wurde und wird nach chromosomalen Abschnitten im menschlichen Genom gesucht, an denen die klinische Diagnose Dyslexie ihr physisches, molekulargenetisches Pendant findet. In verschiedenen Kopplungs- und Assoziationsstudien wurden bislang verschiedene Loci entdeckt, die in der Entwicklung der Dyslexie eine Rolle spielen (siehe dazu Abschnitt 2.3).

Zu den kognitiven Funktionen, die an der Störung der Lese- und Rechtschreibfähigkeit beteiligt sind, zählen das phonologische Bewusstsein, das orthographische Wissen und die Wahrnehmung und Verarbeitung akustischer und visueller Informationen. Das phonologische Bewusstsein setzt sich aus den Fähigkeiten der Lautanalyse und des Lautgedächtnis zusammen. Mehrere Studien haben die Bedeutung des Faktors phonologisches Bewusstsein gezeigt. Bereits Risikokinder – Kinder aus Familien mit bekannter Häufung von LRS - im Vorschulalter zeigen hier eine Einschränkung (Torgesen et al., 1992; Marx et al., 1993), die sich auch bei betroffenen Schulkinder und Erwachsenen mit LRS nachweisen lässt (Schulte-Körne, 2001b).

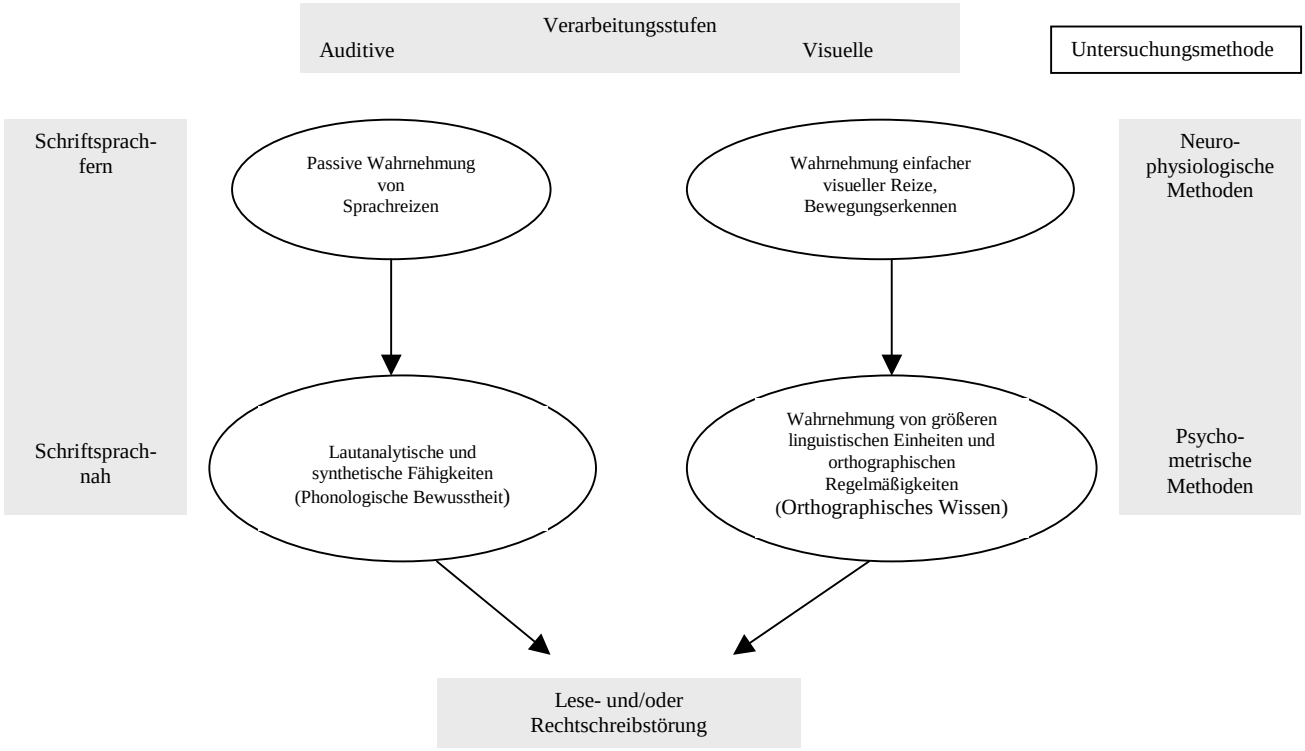

Abbildung 2.2
Auditive und visuelle Verarbeitungsstufen, die ursächlich mit der LRS in Verbindung stehen - nach Remschmidt (2000)

Das orthographische Wissen steht für die Fähigkeit, Regelmäßigkeiten in Buchstabenfolgen zu erkennen und schriftlich wiedergeben zu können. Die für das Lesen und Schreiben wichtigen akustischen und visuellen Wahrnehmungen und Verarbeitungen wurden neurophysiologisch untersucht. Die Abbildung 2.2 zeigt die Ergebnisse dieser Untersuchungen.
Demnach werden die beiden Verarbeitungssysteme wiederum in verschiedene Verarbeitungsstufen unterteilt, welche sich zwar gegenseitig beeinflussen können, die aber jede für sich spezifisch zur Ausbildung der Störung beitragen können (Remschmidt, 2000).
Weiter zeigte sich bei betroffenen Kindern und Erwachsenen sowie Säuglingen betroffener Eltern eine geringere kortikale Aktivierung bei Sprachreizen. Der Einfluss des visuellen Systems und besonders der Okulomotorik – hier z.B. Blicksprünge oder verlängerte Fixationszeiten – ist derzeit noch relativ unklar (Schulte-Körne & Remschmidt, 2003).
Die Rolle der Umweltfaktoren wurde früher wohl überschätzt und scheint tatsächlich eher gering zu sein. Bisher wurde nur eine Zwillingsstudie durchgeführt, die als Umweltfaktoren die Familiengröße, die Herkunft der Eltern, das aktuelle sozialen Umfeld sowie die familiären Lesegewohnheiten untersuchte. Nur 6% der Lese- und 13% der Schreibfähigkeit wurden durch Umweltfaktoren erklärt (Stevenson, 1990).

2.2 Dyslexie als Phänotyp, Wahl des Studiendesign

Bevor die Diagnose LRS gestellt wird, müssen andere Ursachen für Lese- und Rechtschreibprobleme ausgeschlossen werden. So dürfen keine sonstigen organischen, neurologischen oder psychiatrischen Krankheiten als Ursache vorliegen, die Entwicklung und die Intelligenz müssen normal und die Beschulung angemessen sein. Von den möglichen neurologisch-psychiatrischen Ursachen spielt das Hyperkinetisches Syndrom (HKS) eine große Rolle und muss von der LRS abgegrenzt werden (Remschmidt, 2002).
Nach den Kriterien der DSM-IV (A.P.A., 1994) und des ICD-10 muss die erbrachte Leseleistung wesentlich unter dem aufgrund des Alters, des Intelligenzquotient (IQ) und der durch Erziehung erwarteten Ergebnisse liegen. Hierbei sollte die Differenz zwischen IQ und erbrachter Leistung mindestens zwei Standardabweichungen, bei niedrigem IQ 1,5 Standardabweichungen betragen, da erst dann von einer Störung anstatt eines mangelnden Denkvermögens ausgegangen werden kann. Für eine sichere Diagnose muss der IQ höher als 85 sein. Ein weiteres Kriterium der Lesestörung ist nach der DSM-IV, dass sie den

schulischen Verlauf bzw. alle Aktivitäten im täglichen Leben, die mit Lesen zu tun haben, beeinflusst.

Bei Untersuchungen zur LRS müssen diese Aspekte berücksichtigt werden. So wurden für die Studie, in deren Rahmen diese Arbeit entsteht, bei der Probandenauswahl Diskrepanzkriterien für Alter und IQ eingeführt, die Beschulung wurde berücksichtigt, und mögliche kausale Erkrankungen wurden ausgeschlossen. Außerdem mussten die Probanden aus einem deutschsprachigen Elternhaus stammen.

Um eine Untersuchung des möglichen genetischen Hintergrundes der Dyslexie durchzuführen, bedarf es der Klärung, was als Dyslexie bezeichnet werden soll, da ein Phänotyp *Dyslexie* für weitere Analysen benötigt wird. Für die Phänotypisierung gibt es prinzipiell zwei Möglichkeiten: Man kann entweder qualitative oder quantitative Phänotypen verwenden. Qualitativ bedeutet, dass der Phänotyp binär ist, also z.B. betroffen oder nicht betroffen. Dagegen können quantitative Phänotypen auf einer Skala ausgedrückt werden. So könnte sich etwa für den Phänotyp *Lesen* ein Wert zwischen 1 und 10 angeben lassen. Im Rahmen der LRS-Studie werden quantitative Phänotypen verwendet. Sie haben den Vorteil, dass sie in verschiedenen Tests (z.B. Variationen in verschiedenen Altersgruppen) vergleichbar sind, sie sind bei multifaktoriellen multigenen Erkrankungen „Gen-näher“ und haben die statistisch größere Power (Geller & Ziegler, 2000). Darüber hinaus haben quantitative Phänotypen den Vorteil, dass man sie in verschiedenen Studien besser vergleichen kann.

Für die Untersuchung der genetischen Komponente der Dyslexie scheinen erweiterte Stammbäume zunächst viel versprechend. Der überwiegende Nachteil ist aber, dass die entfernten Verwandten der Betroffenen oft schwer oder nicht erreichbar sind. Außerdem existieren für die Diagnose der Dyslexie keine Testvorlagen für Erwachsene, genauer: für Schüler ab der 10. Klasse.

Für die molekulargenetische Untersuchung der Dyslexie sind deshalb Geschwisterpaarstudien unter Verwendung quantitativer Phänotypen die Methode der Wahl. Sie haben meist eine hohe Power, eine einfach durchzuführende Rekrutierung und eine einfache Datenstruktur. Darüber hinaus ist die Bereitschaft zur Teilnahme meist dann höher, wenn nahe statt entfernte Verwandte betroffen sind. Die Geschwisterpaarstudie basiert auf der Idee, dass sich phänotypisch ähnliche Geschwister auch im Genotyp ähneln sollten.

Man unterscheidet drei Studiendesigns für die Geschwisterpaaranalyse:

- *Random sib pair* (RSP)-Design
  (beide Geschwister werden rein zufällig aus der Bevölkerung ausgewählt)

- *Single proband sib pair* (SPSP)-Design
  (eines der beiden Geschwister hat eine extreme phänotypische Auswertung)
- *Extreme sib pair* (ESP)-Design
  (beide Geschwister haben eine extreme phänotypische Ausprägung)

Extreme disconcordant sib pairs-Designs mit Geschwistern, die eine stark unterschiedliche Ausprägung des Merkmals haben, haben die größte Power, um Kopplung zu finden (Risch & Zhang, 1995). Dies ist jedoch häufig dadurch begründet, dass die Kinder verschiedene biologische Eltern haben (*Non-paternity*) (Neale, Neale et al., 2002; Ziegler et al., in press).
Quantitative Phänotypen scheinen nicht einem einfachen biallelischen Mendel'schen Erbgang zu folgen, sondern eher einem oligogenen Modell (Gilger et al., 1994). Für solche Vererbungsmodelle erhöht das ESP-Design nicht immer die Power (Allison et al., 1998). Die Anzahl der Geschwister, mit denen potentiell andere Phänotypen untersucht werden können, würde sich stark reduzieren, wenn ein ESP-Design mit dem am meisten interessierenden Phänotypen angewandt würde. Um mehrere quantitative Phänotypen gleichzeitig zu untersuchen, ist dieser Ansatz also nicht gut geeignet (Ziegler, 1999).
Bei der LRS-Studie wurde ein SPSP *(Single-proband-sib-pair)*-Design verwendet. Ein betroffener Proband (Kind mit phänotypischer Ausprägung LRS), mindestens ein Geschwisterkind und die Eltern wurden erfasst. Hat der Proband einen extrem niedrigen phänotypischen Wert und das zweite Kind am relevanten Genort den gleichen Genotyp, dann sollte es ebenfalls einen erniedrigten phänotypischen Wert aufweisen. Die Begründung für diese Wahl des Studiendesigns als das am besten geeignete ist von Ziegler dargelegt worden (Ziegler, 1999).
Die betroffenen Probanden und die Geschwister wurden mit verschiedenen Tests untersucht. So wurden Intelligenzmessungen, Tests zu Rechtschreib- und Leseleistung sowie psychometrische Tests und EEG-Tests durchgeführt. Es wurde eine Blutprobe entnommen, um den Genotyp bestimmen zu können (Näheres siehe Kapitel 3.3).
Von den Eltern fließt nur der Genotyp in die Auswertung ein. Bei den Studiendesigns zur Geschwisterpaaranalyse wird der Phänotyp der Eltern nicht berücksichtig. Dies ist ein Vorteil, da – wie erwähnt - bei der Dyslexie die Phänotypisierung Erwachsener mit den vorhandenen Tests nicht befriedigend gut möglich ist.

2.3 Bisherige Studien zur Genetik der Dyslexie

Verschiedene Segregationsanalysen zur Dyslexie wurden bisher durchgeführt, um Hinweise auf das zugrunde liegende genetische Modell zu erhalten. Die Ergebnisse sind mitunter widersprüchlich. Lewitter zeigte die Kompatibilität mit einem rezessivem Hauptgen für Familien mit weiblichen Probanden und lehnte alle anderen Modelle ab (Lewitter et al., 1980). Pennington zeigte einen additiven oder dominanten Hauptgen-Effekt (Pennington et al., 1991). Unumstritten ist heute, dass die Dyslexie nicht nach einem einfachen Mendel'schen Modell vererbt wird.

Bisherige Assoziationsstudien und Kopplungsanalysen deuten auf Loci, also Genorte, auf den Chromosomen 1, 2, 3, 6, 7, 11, 13, 15, 18 und dem X-Chromosom. Die Studien sind entweder auf eine bestimmte chromosomale Region gerichtet oder untersuchen das gesamte Genom (Genomscan). Die erste Kopplungsanalyse wurde 1983 von Smith durchgeführt (Smith, 1983). Hier ergab sich der Hinweis auf Kopplung zu Chromosom 15. Spätere Studien wiesen ebenfalls auf Chromosom 15 als mögliches Kandidatenchromosom (Smith, 1991; Grigorenko, 1997; Schulte-Körne, 1998b). Die erste genomweite Untersuchung wurde 1999 von Fagerheim durchgeführt, der eine Kopplung zum Chromosom 2 nachweisen konnte (Fagerheim et al., 1999).

Neuere Studien untersuchten genauer Loci auf dem Chromosom 6 und konnten eine Reihe von Suszeptibilitätsgenen – also Gene, die die Erkrankungswahrscheinlichkeit beeinflussen – identifizieren (Francks et al., 2004; Deffenbacher et al., 2004; Cope et al., 2005; Schumacher et al., 2006).

Tabelle 2.1 gibt eine Übersicht in chronologischer Reihenfolge über die bisher stattgefundenen molekulargenetischen Untersuchungen ohne Anspruch auf Vollständigkeit.

Molekulargenetische Untersuchungen zur Dyslexie

Autoren	Chromosom	Methodik / Design	Phänotyp	Ergebnis	Land (Sprache)
Smith et al., 1983	1	Mehrgenerationenfamilie, psychometrische Tests und anamnestische Angaben, klassische Kopplungsanalyse	Leseschwäche	Hinweis für Kopplung	USA
Smith et al., 1991	6 und 15	18 Mehrgenerationen-familien, psychometrische Tests, sib-pair-Analyse	Leseschwäche	Hinweis für Kopplung und Heterogenität	USA
Rabin et al., 1993	1	9 Mehrgenerationenfamilien	Leseschwäche	Hinweis für Kopplung	USA
Cardon et al., 1994	15 und 6	19 Mehrgenerationen-familien und Zwillingsstichproben, Psychometrische Tests, sib-pair-Analyse (QTL)	Lese-Rechtschreibschwäche (Score aus Wort-Lesen, Leseverständnis und Rechtschreibung)	Hinweis für Kopplung	USA
Grigorenko et al., 1997	15 und 6	6 Mehrgenerationenfamilien, klassische Kopplungsanalyse (modellbasierte und modellfreie Analysen)	5 verschiedene Phänotypen (phonologisches Bewusstsein, phonologisches Dekodieren, Objekt-Benennen, Wort-Lesen, Diskrepanz-kriterium zwischen IQ (gemessen anhand Wortschatz und Lesefähigkeit)	Kopplung des Phänotyps „Phonologisches Bewusstsein“ zu Chromosom 6, Kopplung des Phänotyps „Wort-Lesen“ zu Chromosom 15	USA
Schulte-Körne et al., 1998	15	7 Mehrgenerationenfamilien, psychometrische Verfahren und anamnestische Angaben, klassische Kopplungsanalyse (modellbasierte und modellfreie Analysen)	Rechtschreibstörung	Hinweis für Kopplung	Deutschland

Autoren	Chromosom	Methodik / Design	Phänotyp	Ergebnis	Land (Sprache)
Field et al., 1998	6	79 Familien (davon 30 Mehrgenerationenfamilien), psychometrische Verfahren und anamnestische Angaben, klassische Kopplungsanalyse (modellbasierte und modellfreie Analysen), Assoziationsstudie	Betroffenheit definiert anhand eines Untersucher-Ratings basierend auf verschiedenen Testergebnissen zum phonologischen Bewusstsein, Lesen und Rechtschreibung; bei Erwachsenen zusätzlich anamnestische Angaben	Kein Hinweis für Kopplung	Kanada (Englisch)
Fisher et al., 1999	6	82 Familien, 181 Geschwisterschaften, Psychometrische Tests, sib-pair-Analyse (QTL)	Wort-Lesen, Diskrepanz aus Wort-Lesen und IQ, Orthographisches Wissen, phonologische Dekodierfähigkeit (Nichtwort-Lesen)	Kopplung des Phänotyps Orthographisches Wissen und phonologisches Dekodieren zu Chromosom 6	Großbritannien
Gayan et al., 1999	6	79 Familien, 126 Geschwisterhälften, Psychometrische Tests, sib-pair-Analyse (QTL)	Wort-Lesen, Orthographisches Wissen, phonologisches Dekodieren (Nichtwort-Lesen), phonologisches Bewusstsein (pig-latin, Laute-Streichen)	Kopplung des Phänotyps Orthographisches Wissen und phonologisches Dekodieren zu Chromosom 6	USA
Fagerheim et al., 1999	2p15-16	1 Familie mit autosomal-dominantem Auftreten von Dyslexie, Kopplungsanalyse bei 36 Familienmitgliedern	Lesen	Hinweis für Kopplung, wahrscheinliche Lage auf 2p15-p16 zwischen D2S1337 und D2S2352	Norwegen
Morris et al., 2000	15q	Familienbasierte Assoziation, 101 bzw. 71 Familien	Lesen	Assoziation zu D15S994 D15S214 D15S146	Großbritannien
Petryshen et al., 2000	6p23-p21.3	79 Familien, sib-pair-Analyse (QTL) und variance components analysis	4 Phänotypen (phonologisches Bewusst-sein, phonologisches Dekodieren, Objekt-Benennen, Buchstabieren)	Keine Kopplung nachweisbar, keine Assoziation nachweisbar	Kanada (Englisch)

Autoren	Chromosom	Methodik / Design	Phänotyp	Ergebnis	Land (Sprache)
Petryshen et al., 2001	6q	96 Familien mit mindestens zwei betroffenen Geschwistern, 2-Punkt-modellbasierte Analyse und Mehrpunkt-Kopplungsanalyse	Qualitativer Genotyp (krank, gesund, unklar)	Hinweis für Kopplung D6S254 D6S965 D6S280 D6S251	Kanada (Englisch)
Grigorenko et al., 2001	1p und 6p	Modellbasierte und modellfreie Kopplungsanalyse, 8 erweiterte Familien (mit mindestens vier Betroffenen)	6 Phänotypen (phonologisches Bewusstsein, phonologisches Dekodieren, Objekt-Benennen, Wort-Lesen, Vokabular, Diagnose „lebenslange Dyslexie“)	Hinweise auf Locus 1p und 6p	USA
Nopola-Hemmi et al., 2001	3	140 Familien, Genomscan mit 320 Markern	Neurophysiologische Tests	Kopplung eines Locus auf Chromosom 3	Finnland
Fisher et al., 2002	18	2 Zwillingsstudien (QTL) mit 89 englischen und 119 amerikanischen Familien, Genomscan	Wort-Lesen, orthographisches Wissen, phonologisches Dekodieren (Nichtwort-Lesen), phonologisches Bewusstsein (pig-latin, Laute-Streichen)	Kopplung zu Chromosom 18p11.2	Großbritannien
Kaplan et al., 2002	6p21.3-22	104 Familien, Haseman-Elston- und DeFries-Kopplung (QTL)	11 quantitative Phänotypen (Lesen und Schreiben)	Hinweis für Kopplung	USA
Kaminen et al., 2003	2p11	11 Familien mit 38 dyslektischen Probanden, modellbasierte und modellfreie Kopplungsanalyse	Neurophysiologische Tests	Kopplung zu 2p11 und Verdacht auf möglichen neuen Locus 7q32	Finnland
Nopola-Hemmi et al., 2003	2p11	Nachfolgestudie zur Studie von Kaminen et al. (2003)	Neurophysiologische Tests	Eingrenzung auf chromosomales Gebiet D2S2116-D2S2181	Finnland

Autoren	Chromosom	Methodik / Design	Phänotyp	Ergebnis	Land (Sprache)
Grigorenko et al., 2003	6p	107 Probanden, Kopplungsanalyse	Phonologisches Bewusstsein, phonologisches Dekodieren, Objekt-Benennen, Wort-Lesen	Kopplung zu 6p21.3, zwischen Markern D6S105 und D6S265 und um die Marker D6S109 und D6S1261	USA
Hsiung et al., 2004	11p15.5	100 Familien mit mindestens zwei betroffenen Geschwistern, Kopplungsanalyse	Quantitative Phänotypen	Kopplung vom Dyslexie-Locus (DYX7) an Dopamin-Rezeptor-Region (DRD4) auf Chromosom11p15.5.	Kanada (Englisch)
Tzenova et al., 2004	1p34-p36	100 Familien, Kopplungsanalyse (QTL)	Qualitativer Phänotyp (gesund, krank, unklar) und quantitative Phänotypen (phonologisches Bewusstsein, phonologisches Dekodieren, Buchstabieren, Objekt-Benennen)	Kopplung zu Chromosom 1p34-p36	Kanada (Englisch)
Francks et al., 2004	6p22.2	Familienbasierte Assoziation, 264 englische und 175 amerikanische Familien	Quantitative Phänotypen	Assoziation zu den Genen TTRAP, THEM2 und KIAA0319	Großbritannien und USA
Chapman et al., 2004	15q	111 Familien	Phonologisches Dekodieren, Wort-Lesen	Beweis für Kopplung zu Chromosom 15q	USA
De Kovel et al., 2004	Xq27	Genomscan einer Großfamilie	Wort-Lesen, Nicht-Wort-Lesen, phonologisches Dekodieren	Hinweis auf Kopplung zu Chromosom Xq26-27, Eingrenzung auf Gebiet zwischen den Markern DXS1227 und DXS8091	Holland
Deffenbacher et al., 2004	6p21.3	Haseman-Elston- und DeFries-Kopplung (QTL)	5 quantitative Phänotypen	Kopplung zu 6p21.3 Eingrenzung auf Gebiet zwischen den Markern D6S1597 und D6S1571	USA

Autoren	Chromosom	Methodik / Design	Phänotyp	Ergebnis	Land (Sprache)
Schumacher et al., 2005	18p11-q12	82 Familien, modellfreie Kopplungsanalyse	Quantitative Phänotypen	Kein Hinweis auf Kopplung zu 18p11-q12	Deutschland
Cope et al., 2005	6p22.2	Fall-Kontroll- und Familienbasierte Assoziation	Qualitativer Phänotyp	Assoziation zum Gen KIAA0319	Großbritannien
Igo et al., 2006	13	108 Mehrgenerationen-Familien, Kopplungsanalyse und Segregationsanalyse	Wort-Lesen, phonologisches Dekodieren	Hinweis auf Beeinflussung von Wort-Lesen durch Locus auf Chromosom 13q	USA
Schumacher et al., 2006	6p21-22	239 Familien, modellfreie Kopplungsanalyse	11 quantitative Phänotypen	Hinweis auf DCDC2 als Suszeptibilitätslocus für Dyslexie	Deutschland

Tabelle 2.1

Molekulargenetische Untersuchungen zur Dyslexie

3 Hintergründe zum Dyslexie-Projekt

3.1 Rahmen des Projektes

Der Genomscan findet im Rahmen einer bizentrischen Studie der Kinder- und Jugendpsychiatrien in Marburg und Würzburg statt. An dem Projekt sind die Humangenetik der Universitäten in Würzburg und Bonn, das Institut für Medizinische Biometrie und Statistik in Lübeck (IMBS) und das Max-Planck-Institut für Psychiatrie in München beteiligt.
Für Studie wurden von August 2001 bis April 2004 insgesamt 287 Familien rekrutiert. Alle Teilnehmer der Studie haben vor Durchführung der Tests ihr Einverständnis zur Teilnahme an der Studie gegeben - bei Kindern unter 14 Jahren wurde das Einverständnis durch die Erziehungsberechtigten gegeben. Die Studie wurde durch die Ethikkommissionen der Universitäten in Würzburg und Marburg genehmigt (siehe Ziegler et al., 2005).

3.2 Familien und Studiendesign

Die Datensätze stammen von Familien mit mindestens zwei Kindern, von denen wenigstens eines von einer Lese-Rechtschreibschwäche betroffen ist.
Die Stichproben wurden in den Kliniken für Kinder- und Jugendpsychiatrie und Psychotherapie der Phillips-Universität Marburg und der Bayrischen Julius-Maximilians Universität Würzburg rekrutiert. Von August 2001 bis April 2004 wurden alle Kinder mit Dyslexie in einer der beiden Kliniken untersucht. Es kamen standardisierte und nicht-standardisierte Tests zur Anwendung, und es fand eine Familienanamnese und eine medizinische Anamnese der Kinder statt.
Ein Vorliegen des Hyperkinetischen Syndroms wurde mit Hilfe eines Interviews mit der Mutter überprüft (DIPS[1], basierende auf den Kriterien der ICD-10 für das Hyperkinetische Syndrom), und falls das HKS vorlag, wurden die betroffenen Kinder ausgeschlossen. Dyslexie und Hyperkinetisches Syndrom können eine gemeinsame genetische Pathogenese haben (Remschmidt, 2002; Willcutt et al., 2002), und es sollte ausgeschlossen werden, dass Hyperaktivität und Unaufmerksamkeit die neurophysiologischen und neuropsychologischen Untersuchungen beeinflussen.

[1] DIPS = Diagnostisches Interview bei psychischen Störungen

Neben dem HKS galten als weitere Ausschlusskriterien ein IQ < 85, Störungen des Gehörs und des Sehens, bilinguale Erziehung, ein nicht-deutschsprachiges Elternhaus, Medikationen und ein Alter über 21 Jahre.
Die Resultate der Untersuchungen und Interviews wurden in Marburg gesammelt. Hier wurden sie umcodiert und dann zur weiteren Aufbereitung und statistischen Analyse nach Lübeck ins Institut für Medizinische Biometrie und Statistik weitergegeben. Dort wurden die Daten auch validiert, wobei sie auf Vollständigkeit (Alter, Schulklasse, IQ-Test), Konsistenz bzw. Korrektheit geprüft werden. Die Validierung besteht weiterhin darin zu überprüfen, ob die korrekten Testverfahren für die Rechtschreibtests gemäß dem Untersuchungsmanual durchgeführt worden sind.
Als Studiendesign ist ein SPSP-Design verwendet worden. Ein betroffener Proband (Kind mit phänotypischer Ausprägung LRS), mindestens ein Geschwister und die Eltern wurden erfasst. Die Begründung dieses Designs erfolgte in Kapitel 2.2.

3.3 Phänotypen

Die betroffenen Probanden und die Geschwister wurden mit verschiedenen Tests untersucht. So wurden Tests zur Ermittlung der schulischen Leistungen bei den dyslektischen Kindern und deren Geschwistern durchgeführt. In diesen werden die Bereiche Rechtschreibung, Wortlesen und Nichtwortlesen, *Rapid naming* (schnelles Benennen von Buchstaben, Zahlen, Symbolen und Farben), Phonologie, Rechnen, Pseudohomophone und Reaktionszeit getestet. Darüber hinaus wurden ein Intelligenztest zur Ermittlung der erwarteten Leistung und psychometrische sowie neurophysiologische Tests durchgeführt. Für eine genauere Beschreibung der verschiedenen Tests sei auf den Antrag zur Studie (Remschmidt et al., 2000) und die Publikation von Ziegler et al. verwiesen (2005).
Es wurde eine Blutprobe entnommen, um den Genotyp bestimmen zu können. Von den Eltern fließt nur der Genotyp in die Auswertung ein.
In die Analyse wurden elf aus den Tests resultierende Phänotypen aufgenommen. Die Bezeichnungen der Tests, die Benennungen der daraus resultierenden Ergebnisse und die Bezeichnung der zur Analyse verwendeten Phänotypen sind in Tabelle 3.1 aufgelistet.

Art des Tests	Test-Bezeichnung	Benennung der Testergebnisse	resultierende Phänotypen
Rechtschreibtest	HSP2/3/4/5-9 (Hamburger Schreibprobe) GRT4+ (Grundwortschatz-Rechtschreib-Test), WRT1+/2+ (Weingartener Grundwortschatz Rechtschreib-Test) DRT2/3/4/4+/5 (Diagnostischer Rechtschreib-Test) R-T (Rechtschreibungstest)	rst_t	rst_t
Lese- und Nichtwortlesetest	SLRT (Salzburger Lese- und Rechtschreibtest) Marburger Leselisten	SLRT_HW SLRT_WU WL_Z NWL_Z	 ph_nwl, ph_les
Rapid Naming	Buchstaben Zahlen Symbole Farben	RAP-B-Z RAP-Z-Z RAP-S-Z RAP-F-Z	ph_rap_b ph_rap_z ph_rap_sf
Phonologietest	Pseudowort-Segmentierung, Phonemvertauschung, Restwortbestimmung ab 5. Klasse auch Synthesierung und Wortumkehrung	segm2, segm5,vert2, vert5 rest2, synt5 wort5	 ph_phono
Rechentest	Addieren Multiplizieren Punkte zählen	readd_m remul_m punkt_m	 ph_rechn1 ph_rechn2
Pseudohomophone		pseu2rr, pseu2ff pseu5rr , pseu5ff	 ph_pseuho
Kurzzeitgedächtnis	Zahlennachsprechen	zana_wp	zana_wp

Tabelle 3.1

Durchgeführte Tests und ihre phänotypische Bezeichnung

3.4 Kriterien für Dyslexie

Die Diagnose Dyslexie basierte auf dem Score für Rechtschreiben unter Benutzung der T-Verteilung in der normalen Population.
Um in die Studie aufgenommen zu werden, mussten die folgenden Kriterien erfüllt werden: basierend auf der angenommenen Korrelation zwischen IQ und Buchstabieren von 0.4 (Schulte-Körne, 2001), wurde der erwartete Spelling-Score geschätzt. Ein Kind wurde als betroffen eingestuft, wenn die Diskrepanz zwischen dem erwarteten und dem gemessenen Score wenigstens eine Standardabweichung betrug.

3.5 Genotypisierung

Der Genotypisierung wurde von der isländischen Firma deCode durchgeführt. Auf die Technik der Genotypisierung soll hier nicht eingegangen werden.
Für den Genomscan in der LRS-Studie wurde tri- und tetranukleotide STR-Marker verwendet, die aus von der GDB (GDB) ausgewählt wurden. Die Markerdistanzen und Positionen wurden aus der Marshfield-Datenbank (Marshfield Clinic, 2005) und vom UCSC Genome Browser (UCSC Genome Browser) entnommen.
Insgesamt wurden 533 Marker genotypisiert. Von 1060 Personen aus 253 Familien wurden 1058 erfolgreich genotypisiert. Die zwei Personen, die nicht genotypisiert wurden, hatten dieselben Genotypen wie zwei andere Personen im Datensatz.

4 Fehler in molekulargenetischen Untersuchungen

4.1 Bedarf einer Qualitätssicherung

Bei der Untersuchung von Genotypen in Kopplungsanalysen können grundsätzlich zwei verschiedene Arten von Fehlern in den Daten vorliegen: Stammbaumfehler und Genotypisierungsfehler. Ihr Charakter, ihre möglichen Ursachen und Auswirkungen sollen hier näher erläutert werden.

In praktisch allen Untersuchungen, die mit Genotypisierungsdaten arbeiten, liegen Fehler in diesen Daten vor (Bonin et al., 2004; Pompanon et al., 2005). Wird nicht versucht, diese Fehler aufzuspüren, hat dies für die Ergebnisse der Untersuchungen gravierende Auswirkungen. Dies wird in den folgenden Unterkapiteln dargelegt.

Bevor es um die Fehler selbst geht, sollen die genetischen Marker als die „Träger“ dieser Fehler vorgestellt werden.

Am Ende dieses Kapitels werden verschiedene Ansätze zur Vermeidung von und zur Kontrolle auf Fehler erläutert, deren praktische Umsetzung in den Kapiteln 5 bis 10 näher beschrieben werden soll.

4.2 Genetische Marker

Alle Körperzellen des Menschen (außer den roten Blutzellen) enthalten einen Zellkern, der die individuelle genetische Information auf Chromosomen enthält. Eine Zelle, die 22 Autosomen und eines der beiden Gonosomen[1] enthält, wird als haploid bezeichnet. Die meisten menschlichen Zellen enthalten den doppelten Chromosomensatz mit 2 * 22 = 44 Autosomen und 2 Gonosomen. Sie sind diploid, wobei ein Chromosomensatz von der Mutter und einer vom Vater stammt. Die Chromosomen sind aus Desoxyribonukleinsäure (engl. *Desoxyribonuleic acid*, DNA) und Proteinen aufgebaut. Träger der genetischen Informationen ist die DNA. Detallierte Informationen zum Aufbau der DNA sind z.B. bei Strachan und Read oder Ziegler und König nachzulesen (Strachan & Read, 2004; Ziegler & König, 2006).

[1] Autosom = die Chromosomen 1 bis 22, die Chromosomen, die nicht geschlechtsdeterminierend sind
Gonosom = die Geschlechtschromosomen X und Y

Genetische Marker sind Abschnitte auf der DNA, deren Auftreten über die Generationen verfolgt werden kann. Es können Gene oder nicht codierende repetitive Bereiche auf den Chromosomen sein. Jeder Marker hat eine oder mehrere alternative Ausprägungen, Allele. Eine Person kann maximal zwei verschiedene Allele – eines auf dem mütterlichen und eines auf dem väterlichen Chromosom – aufweisen. Liegen zwei gleiche Allele vor, nennt man dies homozygot, bei zwei verschiedenen Allelen spricht man von heterozygot. Die Lokalisation eines Markers ist bekannt.
Marker, die zur genetischen Analyse zur Verfügung stehen, sind:

- RFLPs,
- Minisatelliten,
- Mikrosatelliten und
- SNPs.

Die RFLPs (*Restriction fragment length polymorphisms*) sind die „genetischen Marker der ersten Stunde“. Botstein et al. schlugen sie 1980 als Marker zur Erstellung genetischer Karten vor (Botstein et al., 1980). Sie sind heute nicht mehr gebräuchlich. Ihre Verarbeitung ist teuer und zeitaufwendig. Sie erwiesen sich für das *Disease mapping* – also den Versuch der Kartierung genetischer Erkrankungen – als unbefriedigend, weil sie als 2-allelige Marker nur eine eingeschränkte Informativität haben und sich eine Schlüsselmeiose häufig als uninformativ erweist.

Minisatelliten oder VNTR (*Variable number of tandem repeats*) Marker wurden 1984 von Nakamura entdeckt (Nakamura et al., 1987). Sie bestehen aus 9-80 Basenpaaren[2], haben viele Allele, eine höhere Heterozygotie als die RFLPs, und die meisten Meiosen sind informativ. Allerdings sind die technischen Aufarbeitungen im Southern-Blot und bei der Hybridisierung problematisch. Minisatelliten haben außerdem den Nachteil, dass sie nicht gleichmäßig über das Genom verteilt sind.

[2] Die Purin-Basen Adenin (A) und Guanin (G) und die Pyrimidinbasen Cytosin (C) und Thymin (T) bilden zusammen mit einem Phosphat und einem Zuckermolekül (Desoxyribose) die Bausteine der DNA, Nukleotide. Jeweils zwei Basen (Adenin und Thymin oder Cytosin und Guanin) können untereinander eine schwache Wasserstoffbrückenbindung eingehen und so ein Basenpaar bilden.

Mikrosatelliten, die auch als STR-Marker (*Short tandem repeats*) bezeichnet werden, wurden erstmals von Tautz und Renz als einfache repetitive DNA-Sequenzen beschrieben (Tautz & Renz, 1984). Sie sind bis heute gebräuchlich. Meist bestehen sie aus $(CA)_n$-Wiederholungen, sind kürzer als Minisatelliten und daher besser in der Polymerase-Kettenreaktion (engl. *Polymerase chain reaction*, PCR) zu amplifizieren. STR-Marker können Di-, Tri- oder Tetranukleotide sein. Sie sind gleichmäßig und in großer Zahl auf das Genom verteilt. Für die praktische Anwendung sollten bevorzugt tri- und tetranukleotide STR-Marker verwendet werden, weil die dinukleotiden STR-Marker schwieriger in der PCR zu handhaben sind und beim Scoring zu Stotterbanden neigen (siehe Abschnitt 4.4) (Weber & Broman, 2001; Ghebranious et al., 2003). Neben technischen Problemen beim Scoring haben STR-Marker die Nachteile, dass sie zu Mutationen neigen und für manche Anwendungen nicht in genügend hoher Dichte vorhanden sind (Ziegler & König, 2006).

Von den klassischen RFLPs leiten sich die diallelischen *Single nucleotide polymorphisms* ab (kurz SNPs, sprich: Snips), die seit Ende der 90er Jahre verwendet werden. Ihre geringere Informativität wird dadurch mehr als ausgeglichen, dass sie in großer Zahl über das gesamte Genom verteilt sind und ihre molekulargenetische Analyse schnell und einfach ist (Wang et al., 1998; Strachan, 1999).

Die Abbildung 4.1 zeigt den Aufbau von Mini- und Mikrosatelliten sowie SNPs. Bei SNPs ist eine einzelne Base ausgetauscht, Mini- und Mikrosatelliten haben Sequenzmotive, die unterschiedlich oft wiederholt sind.

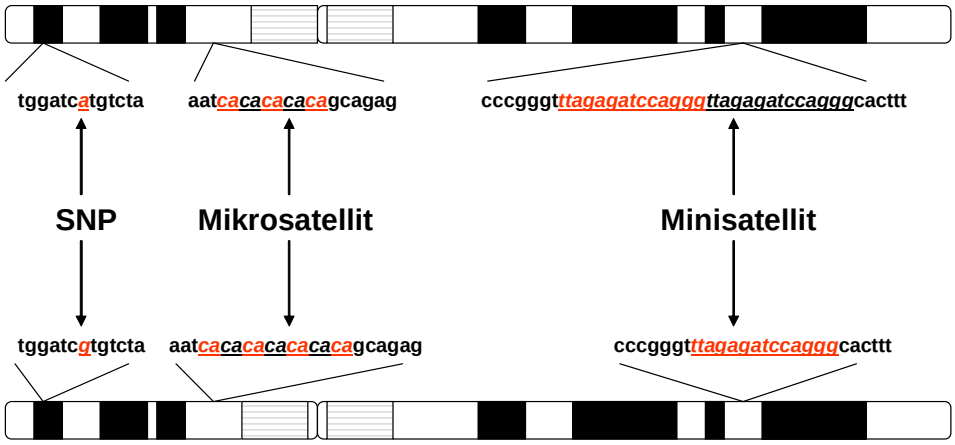

Abbildung 4.1

Genetische Marker – angelehnt an: Cichon et al. (2002)

Für eine detaillierte Beschreibung von STR-Markern und SNPs sei auf die Publikation von Ziegler und König verwiesen (Ziegler & König, 2006).

Die verschiedenen Marker weisen unterschiedliche Raten an Fehlern auf. Die STR-Marker, die auch in der Studie, in deren Rahmen diese Arbeit entsteht, verwendet werden, weisen eine Fehlerhäufigkeit zwischen 0.4% und 3% auf. Dinukleotide Marker weisen höhere Fehleranfälligkeiten auf als tri- und tetranukleotide Marker (Ziegler & König, 2006). Wie es zu Stammbaum- und Genotypisierungsfehlern kommen kann, soll im Folgenden erklärt werden.

4.3 Stammbaumfehler und ihre Ursachen

Stammbaumfehler sind Fehler in der Familienstruktur. Mögliche Ursachen für solche Fehler können unter anderem sein: andere biologische Eltern als angegeben, unbekannte Adoptionen oder vertauschte Proben. Andere biologische Eltern bedeutet in den meisten Fällen andere biologische Väter, also das berühmte Beispiel der „Kuckuckskinder“, die nicht vom bekannten Vater stammen. Es wird geschätzt, dass 3.7% aller Kinder nicht vom vermeintlichen Vater stammen. Bei unentdeckten Kuckuckskindern wissen teilweise nicht nur die Väter sonder auch die Mütter nicht um die wahre Herkunft eines Kindes, weil sie nicht glauben, dass die Kinder tatsächlich noch mit dem vorherigen Partner gezeugt worden sind (Rabbata & Richter-Kuhlmann, 2005; Bellis et al., 2005).

Vor allem bei großen oder verzweigten Familien, in denen die Nachfahren oder entfernten Verwandten nichts von einer Adoption wissen, können Stammbaumfehler auftreten. Vertauschte Proben oder falsche Etikettierungen treten als Fehlerquellen da auf, wo der „Störfaktor“ Mensch arbeitet. Mit strengen Kontrollen und erfahrenem Laborpersonal kann dieser Fehlerquelle bis zu einem gewissen Grade vorgebeugt werden.

Als Folge von Stammbaumfehlern kann die Analyse einen falsch-positiven Nachweis von Kopplung ergeben. Auf der anderen Seite reduzieren Stammbaumfehler die Power der Analyse, und Kopplungen können übersehen werden.

Ein weiterer Fehler, der den Stammbaumfehlern verwandt ist, kann sein, dass einer Person ein falsches Geschlecht zugeordnet wird. Dieser Fehler kann in Überprüfungen des Stammbaumes übersehen werden, wenn zum Beispiel in einer Kernfamilie – Vater, Mutter und Kind – dem Kind ein falsches Geschlecht zugeordnet wird.

4.4 Methoden zum Entdecken von Stammbaumfehlern

Zum Entdecken von Stammbaumfehlern stehen verschiedene Methoden zur Verfügung. Eine Möglichkeit ist die Testung nach McPeek und Sun unter Verwendung der Programme *PREST* und *ALTERTEST* (2002).

PREST steht für *P*edigree *Re*lationship *S*tatistical *T*est. Das Programm überprüft für alle Personen im Stammbaum ihre Relationen zueinander. Das heißt, im Falle einer Kernfamilie mit Vater, Mutter und Kind würde *PREST* überprüfen, ob die Eltern wirklich die Eltern und das Kind wirklich das Kind sein kann. Folgende Relationen können mit dem Programm bestimmt werden: Elter-Kind, Geschwister-Geschwister, Halbgeschwister, Halbgeschwister zu Cousin ersten Grades, Großelter-Enkel, Onkel/Tante, Cousin ersten Grades, Halb-Onkel, Halb-Cousin und nicht verwandt. Das Programm prüft ob die in der Stammbaum-Datei (Näheres dazu siehe Kapitel 5) angegebene Beziehung wahr oder falsch ist.

Wenn *PREST* einen Fehler in einer Relation meldet oder wenn man annimmt, dass zwei Personen im Stammbaum nicht die Relation haben, die für sie angegeben ist, kann das Programm *ALTERTEST* genutzt werden. *ALTERTEST* steht für *Alter*native *Test*. Wie bei *PREST* werden verschiedene statistische Methoden benutzt, um zu berechnen, welche Relation am wahrscheinlichsten ist. Auf die statistischen Hintergründe der beiden Programme sowie ihre Benutzung soll in dieser Arbeit nicht eingegangen werden. Eine genaue Beschreibung findet sich in den Publikationen von McPeek und Sun (2000) sowie Sun et al. (2002).

Geschlechtsangaben in Genotypisierungsdaten sollten mit den Angaben überprüft werden, die bei der Datenerhebung erfasst wurden, um auszuschließen, dass einer Person ein falsches Geschlecht zugeordnet wird.

4.5 Genotypisierungsfehler und ihre Ursachen

Von einem Genotypisierungsfehler spricht man, wenn der in der molekulargenetischen Analyse erhaltene Genotyp nicht mit dem tatsächlichen Genotypen übereinstimmt. Genotypisierungsfehler können verschiedene Ursachen haben.

In jedem Schritt des Prozesses einer molekulargenetischen Untersuchung kann es zu Genotypisierungsfehlern kommen. Die Ursachen können menschlicher oder technischer Natur oder durch die DNA bedingt sein.

Zu Fehlern kann es bei der Probenentnahme, der DNA-Extraktion und Analyse, dem Scoring und der Datenanalyse kommen. Mögliche menschliche Fehlerquellen sind zum Bespiel das Vertauschen von Proben, das unbeabsichtigte Verdoppeln der Proben, Pipettierungsfehler oder Fehler bei der Erfassung, Eingabe und Auswertung der Daten (Bonin et al., 2004).
Bei kontaminiertem DNA-Material und bei geringen oder unvollständigen DNA-Proben treten Genotypisierungsfehler verständlicherweise gehäuft auf (Miller et al., 2002).
Drei wichtige technische Ursachen für Genotypisierungsfehler bei Mikrosatelliten lassen sich unterscheiden:

- Fehler beim Amplifizieren eines der beiden Allele einer Person mit der Folge Allelverlust
- Wiedergabe eines „falschen“ Allels durch die PCR
- Amplifikation einer kontaminierten DNA-Probe

Die wahrscheinlich wichtigste dieser möglichen Ursachen ist der Allelverlust.
Wenn Allele nicht amplifiziert werden, spricht man von Nullallelen. Ein Nullallel kann auf verschiedene Weisen entstehen. Wird bei der DNA-Amplifikation zufällig eines der beiden Allele häufiger amplifiziert als das andere, und handelt es sich um unterschiedliche (also heterozygote) Allele, kann dies zu einer erhöhten Anzahl an Falsch-Homozygoten führen – oder anders gesagt, zu einem Heterozygotenmangel. Eventuell liegt das zweite Allel aber auch außerhalb des beobachteten Bereichs, oder es liegt zwar im untersuchten Abschnitt, kann aber nicht erkannt werden (Chakraborty et al., 1992). Eine Person, die homozygot für ein Nullallel ist, kann als misslungene Beobachtung erscheinen und ausgeschlossen werden (Fisher et al., 2001).

Eine weitere Ursache für Genotypisierungsfehler ist der Verlust des langen Allels / die Dominanz des kurzen Allels, also die bevorzugte Amplifikation von kürzeren Allelen, die ebenfalls in Falsch-Homozygoten resultiert (Van Oosterhout, 2004).

Artefakte in der Elektrophorese können zu Falsch-Heterozygoten oder falscher Zuordnung von Allelen führen. So sind zum Beispiel Stotterbanden Artefakte durch Rückstände, die bei der PCR entstehen und als Verschmierungen die Differenzierung zwischen heterozygot und homozygot erschweren können (Shinde et al., 2003).

Sie sind vor allem ein Problem der dinukleotiden STR-Marker, weniger der tri- und tetranukleotiden STR-Marker (Strachan & Read, 1999). Die Abbildung 4.2 zeigt das PCR-Ergebnis von „normalen“ Banden im Vergleich zu Stotterbanden.

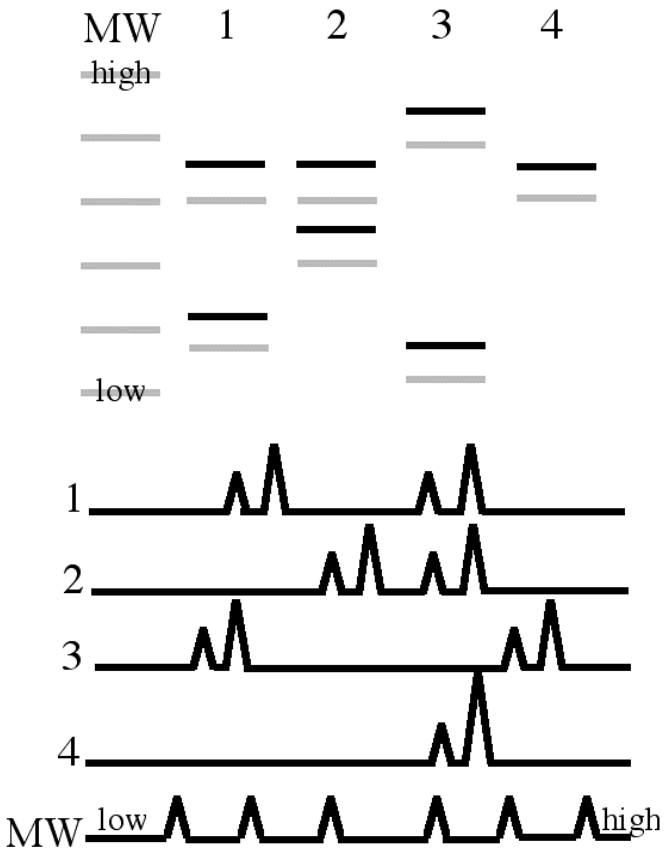

Abbildung 4.2
Beispiel STR-Marker nach Ziegler & König (2006)

Zu sehen sind im oberen Teil der Abbildung vier Datensätze, die mit der Gelelektrophorese dargestellt werden. Die Hauptbanden sind schwarz, die Stotterbanden grau. Im unteren Teil der Abbildung sind die Daten eines Kapillar-Elektrophorese-basierten DNA-Sequenzers zu sehen. Die Peaks repräsentieren verschieden lange/große PCR Produkte. Die Hauptbanden produzieren große, die Stotterbanden kleinere Peaks.

Genotypisierungsfehler können zu der falschen Rekombinanten und Doppelrekombinanten führen. Rekombinanten und Doppelrekombinanten entstehen während des Crossing overs.
Zu Crossing over kommt es während der Meiose. Crossing over bezeichnet das Überkreuzen von zwei Chromatiden (Strachan & Read, 2004). Der Überkreuzungspunkt heißt Chiasma.

Die überkreuzten Chromatiden brechen, und es kommt zum Austausch der entsprechenden DNA-Abschnitte und deren erneutem Zusammenfügen zu Chromosomen. Dadurch vergrößert sich der menschliche Genpool – dieses Phänomen ist damit ein zentraler Prozess der Evolution.
Als Rekombination bezeichnet man ein beobachtetes Crossing over. Zwei Crossing over direkt nebeneinander können nicht beobachtet werden.
Der Begriff Haplotyp bezeichnet einen Block eng benachbarter Allele. Treten zwei Rekombinationen auf einem Haplotyp auf, spricht man von einer Doppelrekombination.

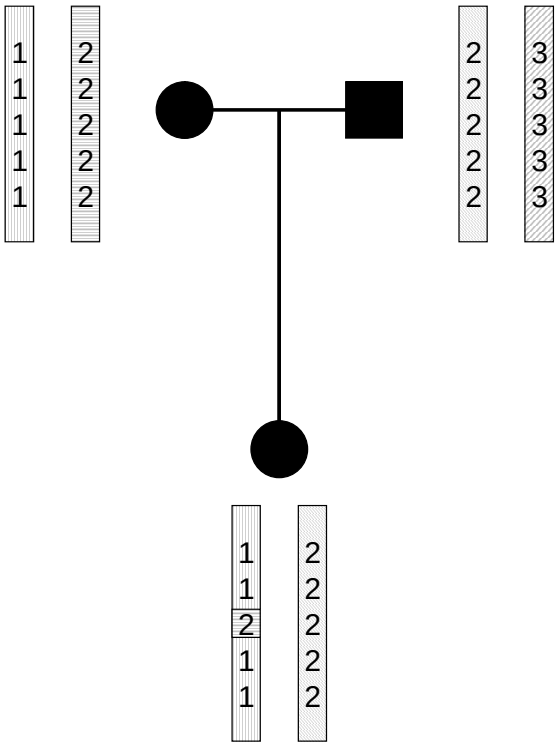

Abbildung 4.3
Stammbaum mit drei Personen; Doppelrekombination beim Kind

Abbildung 4.3 zeigt ein Beispiel für eine Doppelrekombination. Dargestellt ist eine Kernfamilie mit Mutter, Vater und Kind sowie ihren jeweiligen Haplotypen, die für fünf Marker genotypisiert sind.
Das Kind hat einen Haplotypen ohne Rekombinationen vom Vater und einen Haplotypen von der Mutter, der eine Doppelrekombination aufweist. Die erste Rekombination liegt zwischen den Markern 2 und 3. Eine zweite Rekombination liegt zwischen den Markern 3 und 4.

Doppelrekombinationen sind unwahrscheinlich, wenn die beiden Marker sehr dicht zusammen liegen. Rekombinationen treten nämlich nicht unabhängig voneinander auf,

sondern beeinflussen sich gegenseitig. Tritt während der Meiose ein Chiasma auf - also die Überkreuzung der einzelnen Chromatiden der Eltern an einem Punkt – so unterdrückt es in der Regel ein zweites – man spricht von positiver Interferenz.
Eine erhöhte Quote an beobachteten Rekombinationen oder das Auftreten von Doppelrekombinationen können auf Genotypisierungsfehlern beruhen. So kann auch die Doppelrekombination im obigen Beispiel als Genotypisierungsfehler aufgefasst werden. Das Kind könnte am Marker 3 statt homozygot (2/2) auch eher heterozygot wie für die übrigen Marker sein.
Eine erhöhte Quote an Doppelrekombinationen kann aber auch durch wahre Doppelrekombinationen entstehen, weshalb eine Interpretation als Genotypisierungsfehler mit Vorsicht erfolgen sollte.
Falsche Doppelrekombinanten machen etwa ein Viertel der Genotypisierungsfehler bei vollständig typisierten Kernfamilien aus (Douglas et al., 2002). Broman et al. fanden bei der Kartierung der Marshfield Karte bei 0.08% der Genotypen Doppelrekombinanten, die sie größtenteils auf Genotypisierungsfehler zurückführten (Broman et al., 1998).

In Tabelle 4.1 ist eine Übersicht über Ursachen für Genotypisierungsfehler, ihre Entstehungsmechanismen und ihre Auswirkungen zusammengestellt.

Die Konsequenzen von Fehlern sind falsche Rückschlüsse aus den Ergebnisse der Analysen von Genotypisierungsdaten (Pompanon et al., 2005). Fehler können die Rekombinationsfrequenzen erhöhen, damit den Typ 1 Fehler erhöhen und Kartendistanzen verlängern. Die Power der Analyse ist reduziert (Ziegler & König, 2006).
Weiter können Genotypisierungsfehler zu einer falschen Identifikation von Personen führen (Bonin et al., 2004).
Douglas et al. fanden heraus, dass eine einprozentige Rate an Genotypisierungsfehlern zu einem Verlust an Informativität der Kopplungsanalyse von 53% bis 58% führen kann. Eine Fehlerrate von 0.5% verursacht immer noch einen Informativitätsverlust von 28% bis 30% (Douglas et al., 2000). Eine Fehlerrate von nur 3% in analysierten Daten kann bedeutsame Auswirkungen für die Schätzungen zum Kopplungsungleichgewichtes haben (Akey et al., 2001).
Um bei der Durchführung von Kopplungsanalysen sinnvolle Ergebnisse zu erhalten und richtige Schlussfolgerungen ziehen zu können, ist es also wichtig, die Häufigkeit von Genotypisierungsfehlern zu minimieren.

Ursache des Fehlers	Entstehungsmechanismus des Fehlers	Konsequenz des Fehler für den Genotypen
Interaktionen zwischen DNA Molekülen		
An den Marker grenzende DNA-Sequenzen	Keine (oder weniger effiziente) Amplifikation wegen Mutation in der Target-Primer-Sequenz	Nullallele
An den Marker grenzende DNA-Sequenzen	Insertion oder Deletion im amplifizierten Fragment	Längen-Homoplasie von verschiedenen Allelen
An den Marker grenzende DNA-Sequenzen	In heterozygoten Individuen bevorzugte Amplifikation eines Allels	Allelverlust
Qualität des Probenmaterials		
Geringe Qualität oder Quantität der DNA	In Heterozygoten Amplifikation nur eines Allels	Allelverlust
Geringe Qualität oder Quantität der DNA	In Heterozygoten bevorzugte Amplifikation des kürzeren Allels	Dominanz des kurzen Allels
Kontamination des DNA Extraktes	Amplifikation eines kontaminierten Allels	Falsches Allel
Geringe DNA Extrakt Qualität	Keine oder geringere Restriktion oder Amplifikation	Allelverlust
Biochemische Artefakte und technische Probleme		
Geringe Qualität der Reagenzien	Keine oder geringere Restriktion oder Amplifikation	Allelverlust; falsches Allel
Geringe Qualität der Reagenzien	Schlechte Fragmentmarkierung und Erkennung	Allelverlust; falsches Allel
Geringe Reliabilität und Präzision der Ausstattung	z.B. Pipettierungsfehler, Evaporation während der PCR	Allelverlust; falsches Allel
Taq Polymerase Fehler	Verschmierungen	Falsches Allel
Taq Polymerase Fehler	Inkomplett amplifizierte Fragmente	Falsches Allel
Mangel an Spezifität	Amplifikation unspezifischer Produkte wegen *Primer-Annealing* an andere Loci	Falsches Allel
Mangel an Spezifität	Unspezifische Restriktions-Reaktionen	Falsches Allel
Artefakte in der Elektrophorese	Inkonsistenz der Allellängen in verschiedenen Experimenten, Versuchsdurchführungen, Studien	Längen-Homoplasie von verschiedenen Allelen; falsches Allel
Artefakte in der Elektrophorese	Abweichungen der Allellängen durch verschiedene Faktoren, wie z.B. Temperaturunterscheide, Konzentration an PCR-Produkten	Längen-Homoplasie von verschiedenen Allelen; falsches Allel
Menschliche Faktoren		
Probenvertauschen	Vertauschung durch falsche Etikettierungen, verwechselte Röhrchen etc.	Falsche Allele
Fehler in der Versuchsdurchführung	Kontamination mit fremder DNA, Kontamination der DNA Proben untereinander	Falsche Allele
Fehler in der Versuchsdurchführung	Benutzung eines unpassenden Studienprotokolls (z.B. vergessene Reagenz, falsche Primer)	Allelverlust; falsches Allel
Datenverarbeitung	Fehlerhafte Auswertung von Peaks	Falsches Allel
Datenverarbeitung	Fehler oder Verwechslung von Genotypen in der Datenbank	Falsches Allel
Datenverarbeitung	Fehler bei der Datenauswertung (z.B. durch einen Fehler im Analyseprogramm)	Falsches Allel

Tabelle 4.1

Genotypisierungsfehler – angelehnt an: Pompanon et al. (2005)

4.6 Ansätze zur Entdeckung von Genotypisierungsfehlern

Es existieren verschiedene Möglichkeiten die Genotypisierungsdaten auf Fehler zu analysieren.
Einen Teil der Genotypisierungsfehler, aber auch der Stammbaumfehler, kann man durch den Vergleich der Genotypen von Eltern und Kindern entdecken. In diesen so genannten Mendel-Checks wird geprüft, ob die Allele der Kinder mit denen der Eltern kompatibel sind. Kommen bei einem Kind beispielsweise Allele vor, die bei den Eltern nicht gefunden werden können, so liegt ein Fehler vor. Theoretisch könnte auch eine Mutation vorliegen, aber Mutationen sind selten (König mündlich). Das Prinzip der Methode, Mendelfehler zu entdecken, ist in Kapitel 6 beschrieben.

Allerdings können 25% der Genotypisierungsfehler kompatibel mit der Mendelschen Vererbung sein und sind somit mit der Methode des Mendel-Checks nicht zu detektieren (Douglas et al., 2002). Beispiele für Fehler, die kompatibel zur Mendelschen Vererbung sind, sind Fehler, die zu falschen Rekombinationen und Doppelrekombinationen führen. Bei einem Mendel-Check der Familie aus der Abbildung 4.3 würde kein Fehler entdeckt werden können. Das Kind kann ja die Allele von den Eltern geerbt haben.
Um falsche Doppelrekombinationen aufzudecken, können die Haplotypen rekonstruiert und inspiziert werden (Ziegler & König, 2006). Im Kapitel 7 ist dieser Ansatz näher erläutert.

Fehler, wie z.B. Nullallele in den Genotypen können zu Abweichungen vom Hardy-Weinberg Gleichgewicht führen. Das Hardy-Weinberg Gleichgewicht beschreibt die Beziehung zwischen Allelfrequenzen und Genotyphäufigkeiten und besagt, dass diese in aufeinander folgenden Generationen gleich bleiben (Hennig, 2002). Abweichungen vom Hardy-Weinberg Gleichgewicht äußern sich vor allem in einem Homozygotenexzess beziehungsweise in einem Mangel an Heterozygoten.
Das Hardy-Weinberg Gleichgewicht kann genutzt werden, um die Allelfrequenzen abzuschätzen und erhöhte Quoten an Homozygoten zu entdecken. Dieser Ansatz wird im Kapitel 9 weiter verfolgt, wo auch das Hardy-Weinberg Gleichgewicht näher erklärt wird. Wie erhöhte Quoten an Nullallelen entdeckt werden können, wird ebenfalls im Kapitel 9 beschrieben.

5 Qualitätssicherung: Daten und verwendete Dateien

Im Kapitel 4 wurde auf die Ursache von Genotypisierungsfehler eingegangen. Auch erste Ansätze zum Entdecken dieser Fehler wurden diskutiert. In den folgenden Kapiteln sollen einige Methoden zur Qualitätssicherung bei Genotypisierungsdaten konkret vorgestellt werden. Die Benutzung der verwendeten Programme und ihr Output werden erläutert.
Vorher wird in diesem Kapitel auf die wichtigsten Inputdateien eingegangen, die von den Programmen benötigt werden.

Die Qualitätssicherung umfasst sechs Stufen:

- Prüfung auf Mendel-Fehler (Kapitel 6)
- Ausschluss von Doppelrekombinanten (Kapitel 7)
- Kartierung der Marker (Kapitel 8)
- χ^2-Test zum Hardy-Weinberg Gleichgewicht (Kapitel 9.4)
- Monte-Carlo Permutation zum Hardy-Weinberg Gleichgewicht (Kapitel 9.5)
- Test auf Nullallele (Kapitel 9.6)

Um die Genotypen auf Mendel-Fehler zu überprüfen, wird das Programm *Pedcheck,* Version 1.0, (O'Connell & Weeks, 1998) verwendet. Doppelrekombinanten werden mit Hilfe des Programms *Genehunter*, Version 2.1r_5beta, (Kruglyak et al., 1996) ausgeschlossen. Mit dem Programm *Crimap,* Version 2.4, (Green et al., 1990) findet die Kartierung statt. Hierbei werden die Marker auf ihre Reihenfolge geprüft und diese mit vorhandenen Kartierungen verglichen.
Der χ^2-Test auf Abweichungen vom Hardy-Weinberg Gleichgewicht wird mit dem Programm *Pedstats 0.5.4* (Wigginton & Abecasis, 2005) durchgeführt. Die Monte-Carlo Permutation ist in *R* realisiert.
Auf das Vorliegen von Nullallelen wird mit dem Programm *NullAlleleCheck* (Version 0.98, 2004) geprüft.
Außer der Monte-Carlo Simulation, die mit *R* unter *Windows* ausgeführt wird, werden alle Programme auf *Unix* verwendet.

Die im Kapitel 9.7 beschriebene Methode, Vasarely-Abbildungen für die Überprüfung des Hardy-Weinberg Gleichgewichtes zu benutzen, ist als optionale Ergänzung zu den anderen vorgestellten Verfahren zu sehen.

In der Genotypisierung werden für alle Personen für die untersuchten Marker die Allele bestimmt. Die Daten aus der Genotypisierung liegen als *sample.pre* Datei vor.

Je nachdem, mit welchen Programmen man diese Daten untersuchen möchte, benötigt man verschiedene Dateiformate als Input. Zwei Dateiformate sollen hier vorgestellt werden: Die Stammbaum-Datei oder kurz *.ped* Datei, die Informationen zu den untersuchten Personen enthält, und die *.dat* Datei mit Informationen zu den Markern.

5.1 *.ped* Datei

Diese Datei enthält alle Informationen zu den untersuchten Stammbäumen (engl. pedigree).

Jede Familie und jede Person sind eindeutig identifiziert. Der Aufbau der *.ped* Datei geht auf Terwilliger und Ott zurück (Terwilliger & Ott, 1994).

Am Beispiel von zwei Familien aus der LRS-Genotypisierung, die in Abbildung 5.1 dargestellt sind, soll der Aufbau einer *.ped* Datei erläutert werden.

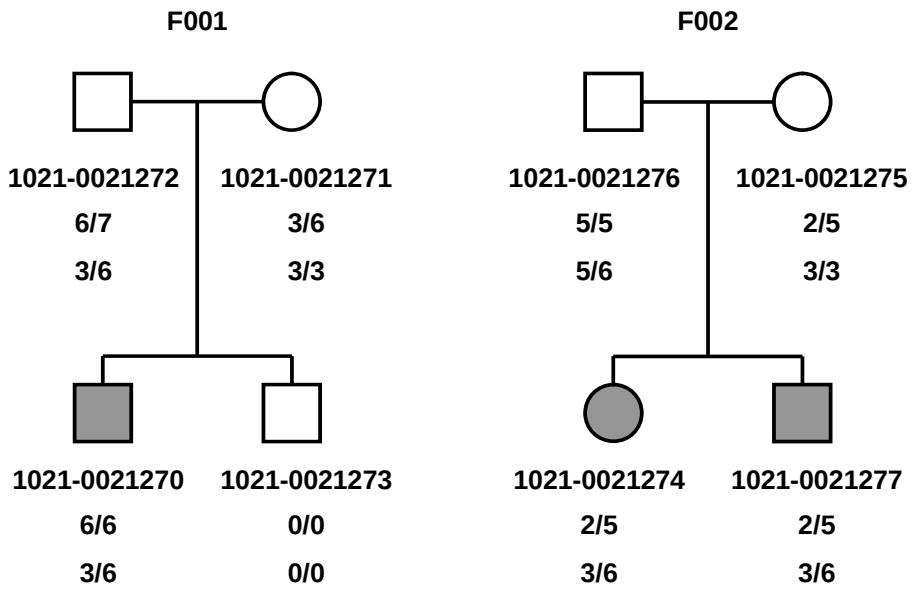

Abbildung 5.1

Stammbäume aus LRS-Genotypisierung

Die Abbildung zeigt die Stammbäume von zwei Kernfamilien. Die Bezeichnungen der Familien (F001 und F002) und der Personen (z.B. 1021-0021272) und die Allele zu zwei Loci sind angeführt (z.B. 6/7 als erster Locus, 3/6 als zweiter Locus bei der Person 1021-0021272). In der Abbildung 5.2 sieht man eine passende *.ped* Datei zu diesen Stammbäumen.

```
F001 1021-0021270 1021-0021272 1021-0021271 1 2 6 6 3 6 38.0029974
F001 1021-0021271 0 0 2 0 3 6 3 3 0
F001 1021-0021272 0 0 1 0 6 7 3 6 0
F001 1021-0021273 1021-0021272 1021-0021271 1 1 0 0 0 0 39.6350529
F002 1021-0021274 1021-0021276 1021-0021275 2 2 2 5 3 6 37.9097491
F002 1021-0021275 0 0 2 0 2 5 3 3 0
F002 1021-0021276 0 0 1 0 5 5 5 6 0
F002 1021-0021277 1021-0021276 1021-0021275 1 2 2 5 3 6 51.7647504
```

Abbildung 5.2
.ped Datei zu den Stammbäumen aus Abbildung 5.1

Jede Zeile in dieser Datei steht für eine Person aus dem Stammbaum. Zur Erklärung wird die erste Zeile betrachtet:

```
F001 1021-0021270 1021-0021272 1021-0021271 1 2 6 6 3 6 38.00299742
```

Zu jeder Familie wird ihre Identifikation (Familien ID – erste Spalte: „F001“) angegeben. Jede Person ist eindeutig durch ihre Identifikation (Person ID – zweite Spalte: „1021-0021270“) und ihre Familien ID bezeichnet. Zu jeder Person wird die Identifikation ihres Vaters und ihrer Mutter (Vater ID – dritte Spalte: „1021-0021272“ – Mutter ID – vierte Spalte: „1021-0021271“) angegeben. Sind Vater oder Mutter unbekannt, wird in die jeweilige Spalte „0“ eingesetzt. Da in diesem Beispiel nur zwei Generationen aufgeführt sind, stehen in den Spalten Vater ID und Mutter ID bei der Elterngeneration Nullen. Man spricht bei den Eltern oder generell bei Personen, denen keine Vorfahren im Stammbaum zugeordnet werden können, von *Foundern*, also Gründern. Personen mit Vorfahren im Stammbaum sind dementsprechend *Non-founder*.
Das Geschlecht der Personen (fünfte Spalte) ist bei männlich mit „1“ – wie für die Person 1021-0021270 –, bei weiblich mit „2“ bezeichnet.
Die sechste Spalte macht eine Angabe zum Betroffenheitsstatus. Der Betroffenheitsstatus macht eine Aussage darüber, ob eine Person von der untersuchten Erkrankung betroffen ist oder nicht. Im Rahmen der LRS-Studie wird diese Aussage nur bei den Kindern getroffen, bei den Eltern ist sie deswegen „0“. Betroffene Kinder sind „2“, nicht betroffene „1“. Die Person 1021-0021270 ist ein betroffenes Kind, deshalb steht bei ihr eine „2“.

Bei jeder Person sind zu jedem Marker die zwei Allele angegeben, die die Person für den Marker aufweist. Ist dies unbekannt oder ist der Marker bei dieser Person nicht genotypisiert, steht eine Null. Im obigen Beispiel sind zwei Marker angegeben. Die Allele des ersten Markers stehen in der siebten und achten, die des zweiten Markers in der neunten und zehnten Spalte; bei der Person 1021-0021270 hat der erste Marker die Allele (6/6) und der zweite (3/6).
Die letzte Spalte gibt einen Wert zum quantitativen Phänotyp an. Da den Eltern in dieser Genotypisierung keine Phänotypen zugewiesen wurden, steht bei ihnen kein Wert („0“). Würde ein qualitativer Phänotyp verwendet, wäre dieser durch den Betroffenheitsstatus hinlänglich beschrieben.
Generell sollten in *.ped* Dateien keine Buchstaben oder Sonderzeichen (wie Bindestriche etc.) stehen, eine Ausnahme kann wie in diesem Beispiel bei der Familien ID und der Personen ID gemacht werden. Die einzelnen Spalten werden durch ein Leerzeichen getrennt.

In der Literatur wird manchmal auch von *.pre* Dateien gesprochen. Oft werden die beiden Begriffe *.ped* Datei und *.pre* Datei nicht klar abgegrenzt. Man kann unter einer *.ped* Datei eine Datei verstehen, die nur Informationen zum Stammbaum und zu den Personen enthält aber nicht zu Markern. In der Literatur zu den Programmen bei Terwilliger und Ott ist eine *post makeped* Datei als *.ped* beschrieben, die vom Programm *Linkage* (Terwilliger & Ott, 1994) benötigt wird. Ihr Inhalt weicht von dem der oben beschrieben *.ped* Datei ab.
Für die Anwendung der Programme zur Qualitätssicherung wird eine *.ped* Datei in der oben beschriebenen Form benötigt.

5.2 *.dat* Datei

Allgemeine Aussagen zum genetischen Modell und zur Anzahl und zu Details über die Marker finden sich in der *.dat* Datei.
Die *.dat* Datei kann vom Benutzer selbst oder durch Programme (z.B. *Preplink* - Funktion von *Linkage*, beschrieben in: Terwilliger & Ott, 1994) erstellt werden. Prinzipiell beschreibt die *.dat* Datei die *.ped* Datei mit Ausnahme der ersten fünf Spalten (Angaben zu den Personen).
Für die Untersuchungen zur Dyslexie sind mehrere Inhalte der *.dat* Datei irrelevant. So wird beispielsweise nicht von einem geschlechts-gebundenen Erbgang ausgegangen. Derartige Angaben fehlen in der Beispieldatei – sie sind gleich „0“ gesetzt.

Die Abbildung 5.3 zeigt das Beispiel einer *.dat* Datei.

```
Zeile
1       5 0 0 5 << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1), PROGRAM
2        0 0.0 0.0 0 << MUT LOCUS, MUT MALE, MUT FEM, HAP FREQ (IF 1)
3           1  2  3  4  5
4       1   2 # disease << AFFECTION, NO. OF ALLELES
5           0.50000  0.50000 << GENE FREQUENCIES
6         1 << NO. OF LIABILITY CLASSES
7             0.0000  0.0000  1.0000 << PENETRANCES
8       3   5 # D1S001 << ALLELE NUMBERS, NO. OF ALLELES
9            0.31250  0.14583  0.10417  0.25000  0.18750 << GENE FREQUENCIES
10      3   3 # D1S002 << ALLELE NUMBERS, NO. OF ALLELES
11           0.39583  0.18750  0.41667 << GENE FREQUENCIES
12      3   4 # D1S003 << ALLELE NUMBERS, NO. OF ALLELES
13           0.37500  0.20833  0.18750  0.22917 << GENE FREQUENCIES
14      0   2  # qtl << QUANTITATIVE, NO. OF ALLELES
15           0.50000  0.50000 << GENE FREQUENCIES
16        1 << NO. OF TRAITS
17                0.000     0.000     0.000 << GENOTYPE MEANS
18                1.000 << VARIANCE - COVARIANCE MATRIX
19                1.000 << MULTIPLIER FOR VARIANCE IN HETEROZYGOTES
20        0 0 << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
21      0.0000  0.0449  0.0449  0.1000 << RECOMBINATION VALUES
22         1 0.10000 0.45000 << REC VARIED, INCREMENT, FINISHING VALUE
```

Abbildung 5.3
Beispiel für eine *.dat* Datei

Im Folgenden sollen die einzelnen Zeilen der Datei näher erläutert werden.
Hinter den Klammern "<<" stehen jeweils Kommentare, die vom Programm ignoriert werden.
In Zeile 1 ist die "5" ist die Anzahl der Marker: ein Krankheitslocus, drei Marker, die genotypisiert wurden, und der QTL (*Quantitative trait locus* – Erläuterung siehe unten). Ein *Risk locus* oder ein geschlechtsgebundener Erbgang liegen nicht vor, deshalb stehen in den entsprechenden Spalten Nullen. Bei einem X-chromosomalem Erbgang stände eine „1" an der dritten Position. Die „5" an der vierten Position beschreibt das Programm, für das die *.dat* Datei verwendet werden soll.
Die zweite Zeile dient der Beschreibung eines Locus, der mit einer spürbaren Frequenz mutiert. Ein solcher Locus liegt im Beispiel nicht vor, deshalb sind alle Angaben „0" bzw. „0.0".
In Zeile 3 ist die Reihenfolge der Loci in der *.ped* Datei angegeben. Der Betroffenheitsstatus wird in der Regel als erster Locus betrachtet, es folgen die drei Marker mit ihren Allelen und der QTL.
Die „1" in der Zeile 4 deutet dem Programm an, dass die nächsten Zeilen den Krankheitslocus beschreiben. In der *.ped* Datei ist dieser Locus als der Betroffenheitsstatus bezeichnet worden.

Die Anzahl der Allele dieses Locus ist zwei: betroffen „2“ oder nicht betroffen „1“. Ist der Betroffenheitsstatus unbekannt, steht eine „0“.

Die folgenden Zeilen 5, 6 und 7 machen Angaben zu den Genfrequenzen, den *liability classes* (Aussagen zu altersabhängigen Penetranzen) und Penetranzen. Die Penetranz ist der Anteil der Personen, die einen Genotypen für eine Erkrankung aufweisen und auch einen Phänotypen ausprägen (Strachan & Read, 2004). Die drei Möglichkeiten, die in der .dat Datei beschrieben werden, sind homozygot normal, heterozygot und homozygot betroffen. In diesem Beispiel wird davon ausgegangen, dass homozygot Betroffene in allen Fällen (also in 100% deshalb die „1.0000“) erkranken.

Die Zeilen 8 bis 13 beschreiben die genotypisierten Marker – je zwei Zeilen gehören zu einem Marker, also die Zeilen 8 und 9 zum ersten, 10 und 11 zum zweiten usw. Die „3“ in Zeile 8 zeigt dem Programm, dass es sich um Allel-Marker handelt. Die „5“ in Zeile 8 gibt die Anzahl der Allele an. Dahinter steht jeweils der Markername, z.B. „D1S001“. In der Zeile 9 sind die Allelfrequenzen angegeben. Die Allelfrequenzen sind wichtig, wenn durch ein Programm mögliche Genotypen rekonstruiert werden sollen (siehe Kapitel 6).

Die folgenden Zeilen 14 bis 19 machen Aussagen zum QTL. Dieser Locus ist der Ort auf dem untersuchten Chromosom, den man mit Hilfe der quantitativen Variable (Phänotyp) identifizieren möchte (Strachan & Read, 2004). Die „0“ in der Zeile 14 sagt dem Programm, dass der Locus eine quantitative Variable ist, und gibt die Anzahl der Allele wieder. Es folgen Angaben zu den Genfrequenzen, zur Anzahl der Variablen und weitere Informationen zum QTL.

Zeile 20 macht Angaben zu Unterschieden in den Rekombinationsraten von Männern und Frauen und Angaben zu Interferenzen. In diesem Beispiel werden keine solchen Angaben gemacht, die Spalten enthalten Nullen.

Die Zeile 21 enthält Angaben zu den Rekombinationsfrequenzen zwischen je zwei benachbarten Markern. Da in dieser Beispieldatei fünf Marker vorkommen, sind vier Rekombinationsfrequenzen aufgeführt. Der erste Wert ist die Rekombinationsfrequenz zwischen dem Krankheitslocus und dem ersten der genotypisierten Marker (D1S001). Der zweite Wert ist die Rekombinationsfrequenz zwischen D1S001 und dem zweiten Marker (D1S002) usw.

Die letzte Zeile, Zeile 22 im Beispiel, enthält die Nummer der Rekombinationsfrequenz, die variiert werden soll, mit welcher Zunahme variiert werden soll und bis zu welchem Endwert.

5.3 Weitere Dateien

Die Programme benötigen zum Teil weitere Inputdateien. So benötigt *Crimap* verschiedene Inputdateien, die im entsprechenden Abschnitt beschrieben werden.

6 Mendel-Fehler

6.1 Prüfung auf Mendel-Fehler

Eine Möglichkeit, Fehler in den Daten zu entdecken, besteht darin, die Familienstruktur zu nutzen. So kann geprüft werden, ob die angegebenen Genotypen mit der Mendel'schen Vererbung vereinbar sind. Unter anderem werden die erhobenen Daten hinsichtlich ihrer Plausibilität zueinander überprüft. Mendel-Checks erfordern die Genotypen von Verwandten und sind somit nur in Familien-basierten nicht aber in Populations-basierten Studien möglich.
Das Programm *Pedcheck* von O'Connell und Weeks sucht für Autosomen und Gonosomen nach Fehlern in Mendel'schen Erbgängen (O'Connell & Weeks, 1998).
Als Input-Format benötigt *Pedcheck* die Daten der Familien. Hierzu werden eine *.ped* und eine *.dat* Datei wie oben beschrieben benötigt.

Das Programm prüft die Genotypen auf Unstimmigkeiten und meldet Fehler, wenn die folgenden Fälle vorliegen:

- Die Allele eines Kindes und der Eltern sind nicht kompatibel, das heißt, sie entsprechen nicht den Mendel'schen Regeln
- Das Kind ist zwar jeweils kompatibel zu Mutter und Vater einzeln betrachtet, nicht jedoch wenn beide gleichzeitig betrachtet worden.
- Es gibt mehr als vier verschiedene Allele in einer Kernfamilie.
- Es gibt mehr als drei verschiedene Allele bei Geschwistern, wenn ein Kind homozygot für die Allele ist.
- Es gibt mehr als zwei verschiedene Allele bei Geschwistern, wenn zwei Geschwister verschieden homozygot sind.
- Ein Allel liegt außerhalb einer festgelegten Menge von Allelen.
- An einem X-chromosomalen Marker ist eine männliche Person nicht homozygot.
- Eine Person hat in einem als autosomal definierten System nur ein definiertes Allel.

Im Folgenden sollen die Überprüfungen die *Pedcheck* durchführt, an einigen Beispielstammbäumen näher erläutert werden.
Das Programm durchläuft einen fünfstufigen Algorithmus. Die nullte Stufe prüft zunächst, ob alle Personen im Stammbaum verbunden sind.

Auf der ersten Stufe prüft *Pedcheck* die Genotypen der Individuen auf Unstimmigkeiten, so auf Inkompatibilität der Genotypen von Kind und Eltern und auf abweichende Allelzahlen bei Geschwistern. Ein Fehler läge beispielsweise vor, wenn Geschwister mehr als vier verschiedene Allele aufweisen.
Ein derartiger Fehler liegt in Abbildung 6.1 vor.

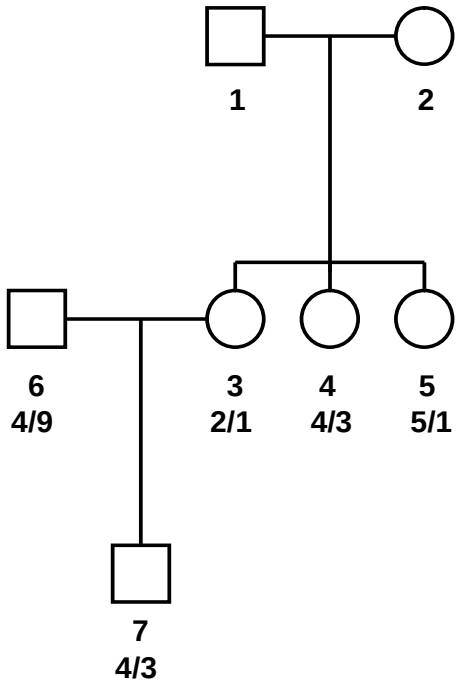

Abbildung 6.1
Stammbaum mit fehlerhaften Allelen

Dargestellt ist ein Stammbaum mit drei Generationen. Die Personen 3-7 sind genotypisiert. Die drei Geschwister – die Personen 3, 4 und 5 - haben die Allele 1, 2, 3, 4 und 5. Somit liegen bei ihnen fünf verschiedene Allele vor. Zwei biologische Eltern können jedoch nur jeweils höchstens zwei verschiedene Allele vererben, so dass Geschwister also höchstens vier verschiedene Allele aufweisen können. Bei einem homozygoten Kind, also etwa mit dem Genotyp (1/1), müssen die Eltern mindestens ein identisches Allel besitzen, also zum Beispiel die Genotypen (1/2) und (1/3) aufweisen. Wenn in dieser Situation unter den Geschwistern mehr als drei verschiedene Allele auftreten, deutet dies ebenfalls auf Fehler hin.
Ein Fehler bezüglich Kompatibilität zwischen den Genotypen von Eltern und Kind liegt in der Kernfamilie mit den Eltern Person 6 und Person 3 und dem Kind Person 7 vor. Das Kind weist den Genotyp (4/3) auf. Die Eltern haben die Genotypen (4/9) – der Vater, Person 6 –

und (2/1) – die Mutter, Person 3. Das Kind ist nicht kompatibel zur Mutter. Es kann nicht das Allel 3 haben, wenn die Personen 3 und 6 die biologischen Eltern sein sollen.

Die Befehlszeile zur ersten Stufe von *Pedcheck* lautet:

```
pedcheck -p pedtest1.ped -d pedtest1.dat
```

Nach dem Befehl "`pedcheck`" werden dem Programm mit "`-p`" und "`-d`" die zu analysierende *.ped* Datei und eine zugehörige *.dat* Datei übergeben.
Die Abbildung 6.2 zeigt das Ausgabe-Format der Fehler, die *Pedcheck* ermittelt, wenn die Daten dieser Familie geprüft werden.

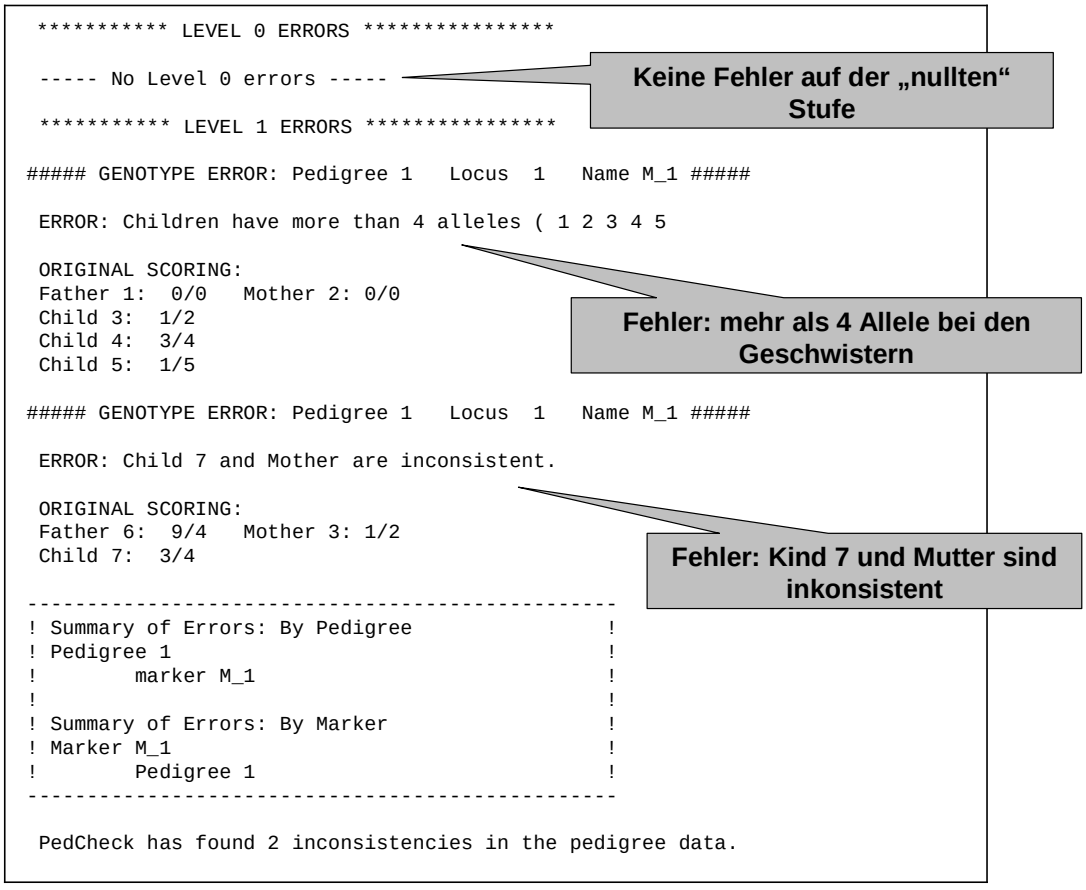

Abbildung 6.2
Bildschirmausgabe zu Stammbaum mit fehlerhaften Allelen

In diesem Stammbaum liegt die Ursache der Fehler wahrscheinlich bei der Person 3. Nimmt man an, dass bei ihr ein Genotypisierungsfehler vorliegt und sie anstelle der Allele (2/1) die Allele (3/1) hätte, wäre der Stammbaum fehlerfrei.

Wenn ein Stammbaum Männer enthält, die nicht hemizygot für X-chromosomale Marker sind - also die Gonosomen XX sind statt XY - erkennt Pedcheck dies auch als Fehler.
Ebenso führen die Allele zu Fehlermeldungen, die außerhalb der in der *.dat* Datei festgelegten Menge von Allelen liegen. Außerdem erfolgt eine Überprüfung, ob alle Personen vollständig typisiert sind. So würde eine Person, bei der nur ein Allel – etwa (1/0) – eingetragen ist, als halbtypisierte Person zu einer Fehlermeldung führen.

Auf der zweiten Stufe führt *Pedcheck* einen Lange-Goradia Algorithmus durch. Dieser dient der genaueren Kontrolle der Allele von Eltern und Kindern. Zur vollständigen Beschreibung des Lange-Goradia Algorithmus sei auf die Publikation von O'Connell und Weeks verwiesen (O'Connell & Weeks, 1999).
Anschließend werden die Kernfamilien betrachtet, also Mutter-Vater-Kind/Kinder. Zu den Genotypen der Eltern werden die jeweils möglichen zygoten Genotypen bestimmt, also z.B. bei einem Genotyp der Mutter (1/2) und des Vater (2/3) ergeben sich die zygoten Genotypen 1, 2 oder 3. Wenn jedes Kind einen oder mehrere dieser Genotypen aufweist, werden die elterlichen Genotypen und die passenden Genotypen der Kinder beibehalten. Wenn ein Kind keine zu den zygoten passenden Genotypen besitzt, werden keine der Genotypen beibehalten.
Für jede Person werden diese Schritte solange durchgeführt, bis keine weiteren Genotypen ausgeschlossen werden können.

In Abbildung 6.3 ist ein Stammbaum mit drei Generationen und zwei Kernfamilien dargestellt. Vier der fünf Personen sind genotypisiert. Die Person 3, Tochter von den Personen 1 und 2 und Mutter von Person 5, ist nicht genotypisiert.

Bei diesem Beispiel sind die Kernfamilien für sich genommen möglich. Insgesamt aber tritt Inkonsistenz auf. Damit der Enkel - Person 5 - die Allele (3/2) haben kann, müsste seine Mutter - Person 3 - wenigstens ein Allel 2 haben. Ihre Eltern haben aber nur die Allele 1, 3 und 4 und können nur diese an ihre Tochter weitergegeben haben. Man erkennt den Fehler im Stammbaum also erst, wenn man ihn über alle 3 Generationen betrachtet.

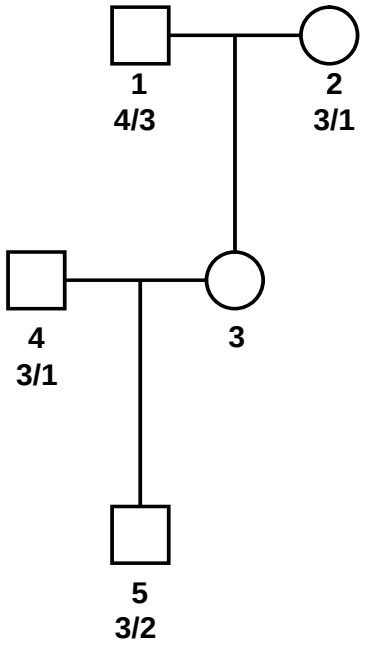

Abbildung 6.3

Stammbaum mit Fehler auf der zweiten Stufe

```
 ********** LEVEL 0 ERRORS ***************

 ----- No Level 0 errors -----

 ********** LEVEL 2 ERRORS ***************

##### GENOTYPE ERROR: Pedigree 1   Locus  1    Name M_1 #####

ORDERED GENOTYPE LISTS: Any allele greater than 4 is set_recoded.
(T) Father 4:    3|1  1|3  (2)
(U) Mother 3:    5|5  3|5  5|3  3|3  (4)

(T) Child 5:    3|2  2|3  (2)

-------------------------------------------------
! Summary of Errors: By Pedigree               !
! Pedigree 1                                    !
!        marker M_1                             !
!                                               !
! Summary of Errors: By Marker                  !
! Marker M_1                                    !
!        Pedigree 1                             !
-------------------------------------------------

 PedCheck has found 1 inconsistency in the pedigree data.
```

Fehler bei den (unbekannten) Allelen der Mutter

Abbildung 6.4

Bildschirmausgabe zu Stammbaum mit Fehlern auf der zweiten Stufe

Zur Durchführung einer Prüfung auf der zweiten Stufe von *Pedcheck* dient der Befehl:

```
pedcheck -2 -p pedtest2.ped -d pedtest2.dat
```

Die Abbildung 6.4 zeigt die Bildschirmausgabe für die Fehler dieses Beispiels.

In der dritten Stufe bestimmt das Programm so genannte „kritische Genotypen“. Darunter werden solche Personen verstanden, deren Elimination aus dem Stammbaum Unstimmigkeiten in demselben aufhebt. Elimination heißt, dass die Personen als unbekannt (0/0) aufgefasst werden. Das führt natürlich zu einem Verlust der Power, aber Fehler werden damit ausgeschlossen.
In der Abbildung 6.5 ist ein Stammbaum, der das Problem der kritischen Genotypen verdeutlichen soll, dargestellt. Die Kernfamilie besteht aus Vater und Mutter – Personen 1 und 2 – und drei Kindern – Personen 3, 4 und 5. Der Vater ist nicht genotypisiert.
Die Mutter ist homozygot mit den Allelen (2/2). Unabhängig von den Allelen des Vaters kann ein Kind der homozygoten Mutter nicht zwei von der Mutter differierende Allele aufweisen. Der Sohn - Person 4 - hat aber die Allele (1/1).
Würde man entweder die Mutter oder den Sohn auf unbekannt (0/0) typisieren, gäbe es ein keine Unstimmigkeiten in diesem Familienstammbaum.

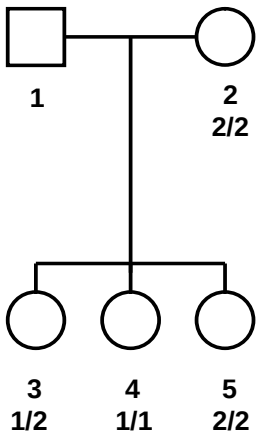

Abbildung 6.5
Stammbaum mit „kritischen Genotypen“

Die Befehlszeile für die dritte Stufe der *Pedcheck* Kontrolle lautet:

```
pedcheck -3 -p pedtest3.ped -d pedtest3.dat
```

Die Abbildung 6.6 zeigt die Fehler-Ausgabe für den Stammbaum mit den kritischen Genotypen. Das Programm empfiehlt, die Person 2 oder die Person 4 auf „0 0“ zu setzen, um im Stammbaum Konsistenz zu erhalten.

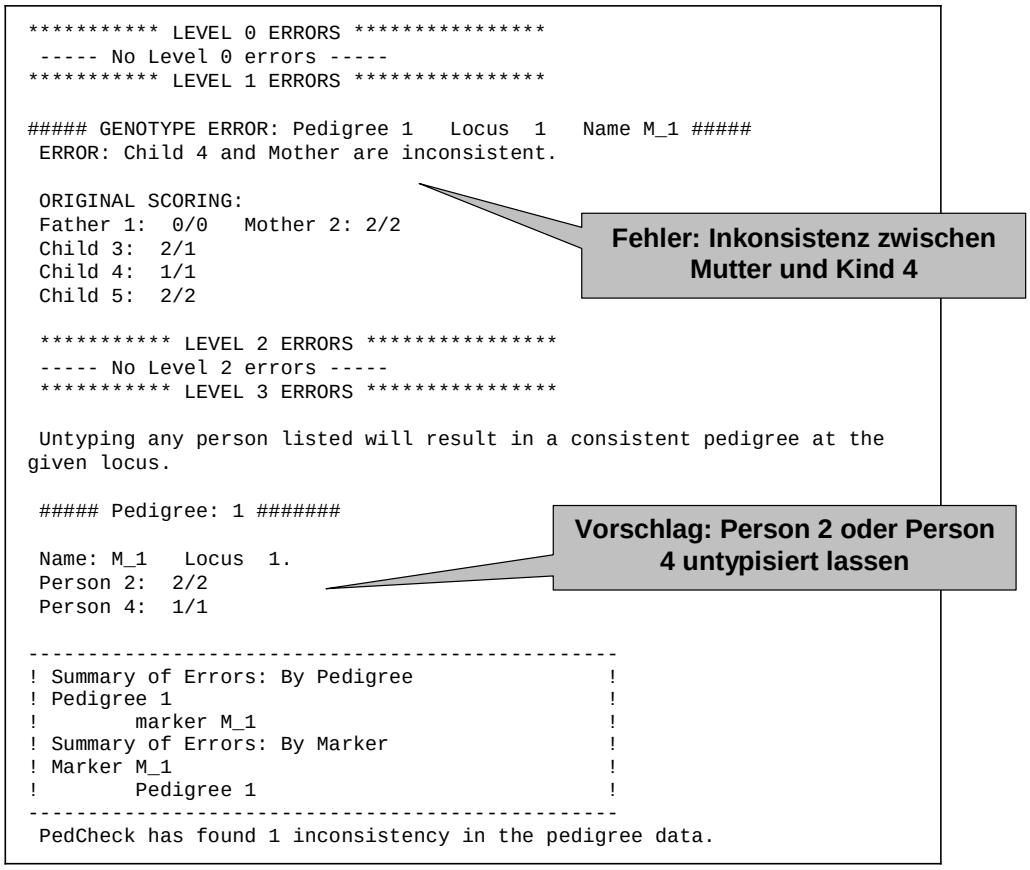

Abbildung 6.6
Bildschirmausgabe zu Stammbaum mit „kritischen Genotypen“

Die vierte Stufe des Programms bestimmt alternative Typisierungen kritischer Genotypen. Für die Familie in Abbildung 3 wären alternative Genotypen für die Mutter der Genotyp (1/2) oder für den Sohn (1/2) und (2/2).
Analog zur vorherigen Kontrolle lautet der Befehl zur vierten Stufe:

```
pedcheck -4 -p pedtest3.ped -d pedtest3.dat
```

Die Abbildung 6.7 zeigt die Bildschirmausgabe für die vierte Stufe.

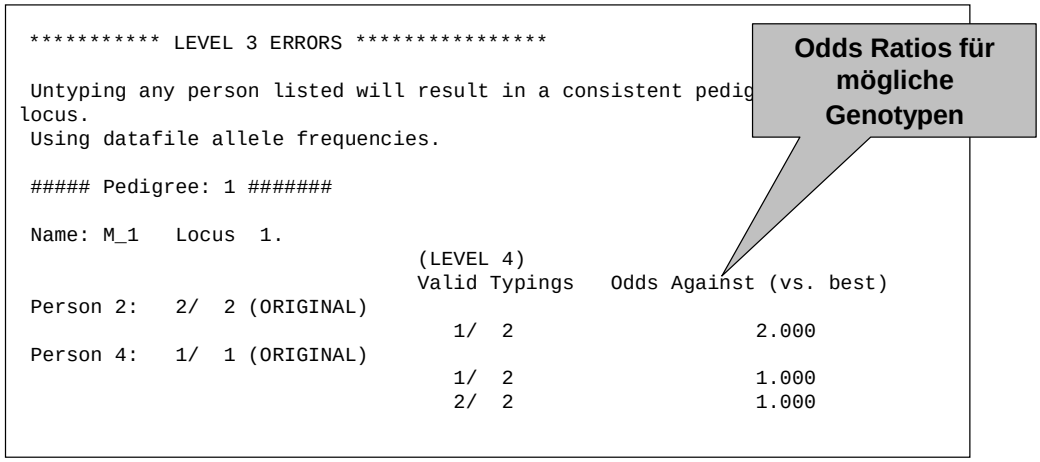

Abbildung 6.7
Ausgabe mit Odds

Pedcheck berechnet auf der Grundlage der Allelfrequenzen, die in der *.dat* Datei stehen, Odds Ratios, mit deren Hilfe man entscheiden kann, welche Individuen am wahrscheinlichsten fehlerhaft typisiert sind oder nicht in den Stammbaum gehören (z.B. Problem der unerkannt nicht-biologischen Väter).
Das Programm führt die dritte und vierte Stufe nur aus, wenn die erste und zweite Stufe fehlerfrei sind.

Das Programm *Pedcheck* ist im Internet erhältlich (Version 1.0, 2000).

6.2 Konsequenz von entdeckten Mendel-Fehlern

Die Genotypisierungsdaten werden wie beschrieben mit *Pedcheck* kontrolliert.
Die in dieser Überprüfung als fehlerhaft erkannten Genotypen werden in der *ped.* Datei von Hand auf „0 0“ gesetzt. Danach können die Daten der modifizierten *.ped* Datei erneut mit dem Programm *Pedcheck* getestet werden.
Finden sich in der Prüfung mit Pedcheck keine Fehler mehr in den Daten, kann als nächste Stufe der Qualitätssicherung die Kontrolle auf Doppelrekombinationen wie im anschließenden Kapitel beschrieben erfolgen.

7 Doppelrekombinationen

7.1 Suche nach Doppelrekombinationen

Wie im Kapitel 4.5 beschrieben, kann eine erhöhte Rate an Rekombinationen oder Doppelrekombinationen auf Genotypisierungsfehler hinweisen.

Ein Beispiel für eine Doppelrekombination die auf einem Genotypisierungsfehler beruht, wurde im Kapitel 4.5 beschrieben. Der Genotypisierungsfehler liegt in diesem Beispiel (siehe Abbildung 4.3) beim Kind, wo auch die Doppelrekombination zu sehen ist.

Eine weitere Möglichkeit, wie es zu falschen Doppelrekombinationen kommen kann, ist als so genannter *phase error* von Broman et al. beschrieben (Broman et al., 1998). Diese Art von Fehler liegt vor, wenn alle oder mehrere Kinder einer Familie an derselben Lokalisation und auf demselben Haplotyp – entweder dem maternalen oder dem paternalen – eine Doppelrekombination zeigen. Die Abbildung 7.1 zeigt so einen Fall.

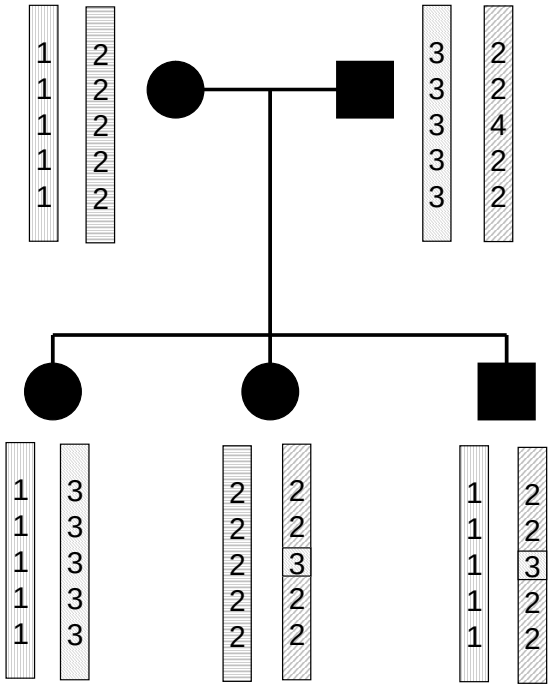

Abbildung 7.1
Familie mit zwei Doppelrekombinationen

In dem Beispiel haben die Eltern am dritten genotypisierten Marker die Genotypen (1/2) und (3/4). Zwei der drei Kinder weisen an diesem Marker eine Doppelrekombination auf. Sie haben die Genotypen (1/3) oder (2/3). Man kann in diesem Fall auf einen Genotypisierungsfehler beim Vater rückfolgern. Statt des beobachteten Genotyps (3/4) könnte

er in Wahrheit den Genotypen (3/3) haben. Dann würden die beiden Kinder auch keine Doppelrekombinationen aufweisen. In diesem Beispiel liegt also der Genotypisierungsfehler gar nicht in der Generation, wo die Doppelrekombinationen auftreten, sondern in der Elterngeneration.
Im Folgenden soll beschrieben werden, wie Doppelrekombinanten gefunden werden können. Aus den beiden Beispielen von oben ergibt sich eine wichtige Konsequenz für entdeckte Doppelrekombinationen. Da man nicht immer sicher sagen kann, wo die Ursache für eine Doppelrekombination wirklich liegt, wird im Falle einer entdeckten Doppelrekombination die gesamte Familie für den betroffenen Marker als (0/0) genotypisiert.

Für den Ausschluss von Doppelrekombinationen wird das Programm *Genehunter*, Version 2.1 5r, verwendet (Kruglyak et al., 1996).
Als Input benötigt *Genehunter* eine *.ped* Datei und eine *.dat* Datei (siehe Kapitel 5).
Es werden die *haplotype* und die *postscript out* Optionen eingeschaltet und die Marker bzw. die *.dat* Datei mit den Informationen zu den Markern mit dem Befehl *load marker* eingelesen. Mit der Funktion *scan pedigree* werden die *.ped* Datei eingelesen und für jede Person in allen Stammbäumen der wahrscheinlichste Haplotyp bestimmt.

```
**************************************************************************
*                                                                        *
*                GENEHUNTER - Complete Linkage Analysis                  *
*                         (version 2.1_r5 beta)                          *
*                                                                        *
**************************************************************************

Type 'help' or '?' for help.
Can't find help file - detailed help information is not available.
See installation instructions for details.

npl:1> ps on
Postscript output is now 'on'

npl:2> haplo on
Haplotype output is now 'on'

npl:3> load marker chr20_nwl.dat
Parsing Linkage marker data file...
14 markers read (last one = D20S171)

npl:4> scan pedigree chr20_nwl.ped

analyzing pedigree F001...
```

Abbildung 7.2
Bildschirmansicht von *Genehunter*

Die Bildschirmansicht von Genehunter mit den Schritten der Befehlseingaben ist in Abbildung 7.2 zu sehen.

Die Haplotypen werden in einer *postscript* Datei gespeichert.
Der Ausschluss von Doppelrekombinanten erfolgt dann „per Hand“, indem bei jeder Familie für jedes Chromosom nach Doppelrekombinationen gesucht wird.

Die Abbildung 7.3 zeigt eine Haplotypenanalyse für eine Familie aus der LRS-Genotypisierung für das Chromosom 20.

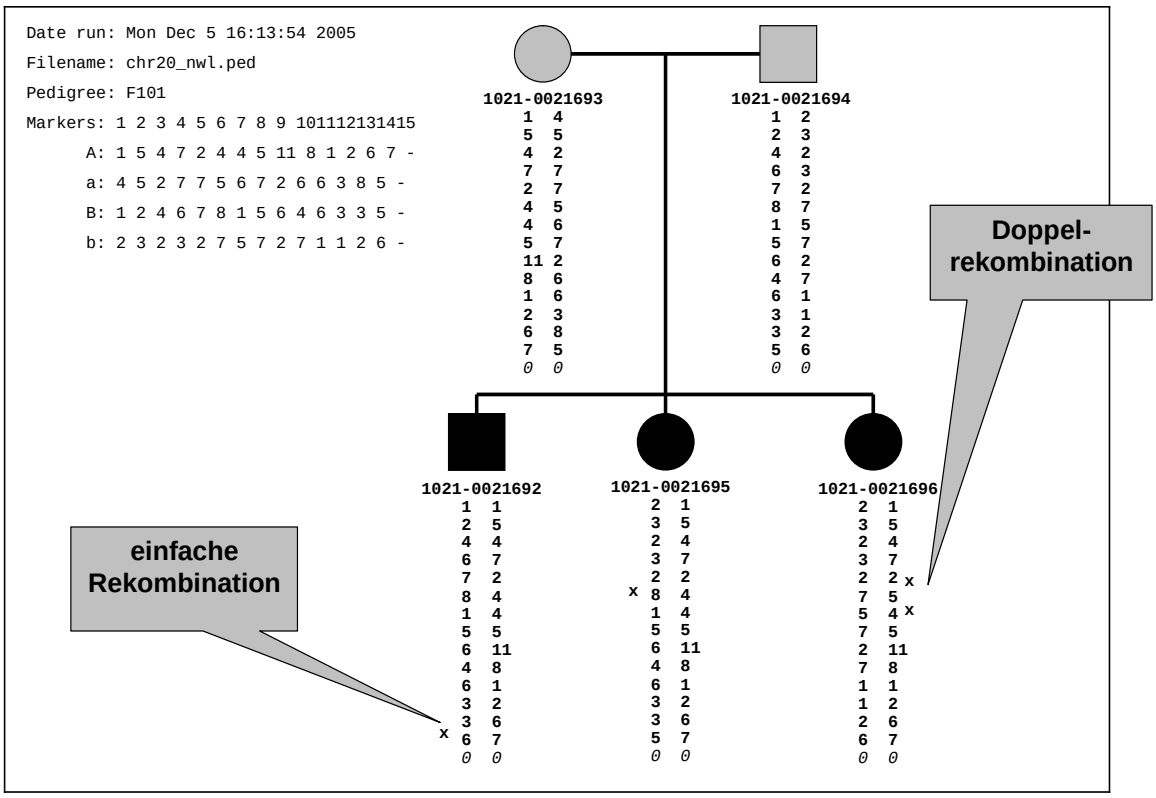

Abbildung 7.3
Haplotypen mit Doppelrekombinante bei einer Familie auf Chromosom 20

Dargestellt sind zum einen nur die Original-Haplotypen der Eltern (A, a, B, b) und der Stammbaum mit den Haplotypen für alle Personen. Rekombinationen sind mit einem „x“ gekennzeichnet. Allele, die fett gedruckt sind, sind die beobachteten Allele. Allele, die dünn und kursiv dargestellt sind, wurden von *Genehunter* bestimmt.
In diesem Beispiel liegt eine Doppelrekombination bei einem Haplotyp der Person 1021-0021696 beim sechsten Marker vor.

Bei zwei so nah beieinander liegenden Rekombinationen wie hier ist die Wahrscheinlichkeit, dass es sich um einen Genotypisierungsfehler handelt, groß – siehe Abbildung 7.4.

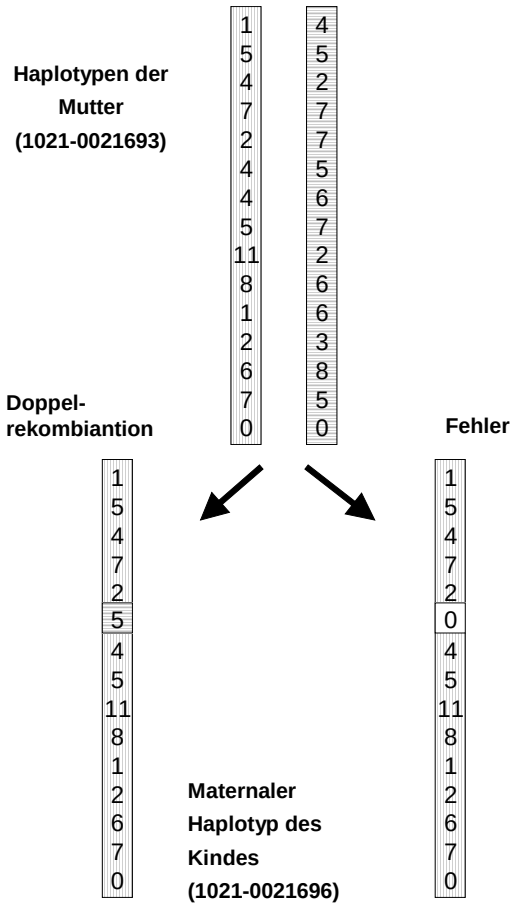

Abbildung 7.4
Doppelrekombination oder Genotypisierungsfehler?

7.2 Ausschluss von Daten mit Doppelrekombinationen

Genotypisierungsfehler können zu falschen Doppelrekombinationen führen. Setzt man die Allele des betroffenen Markers hier auf „0“, ist die Doppelrekombination beseitigt.
So wie in dem obigen Beispiel werden alle Familien für alle Chromosomen inspiziert. Findet man Doppelrekombination, werden die betroffenen Marker in der *.ped* Datei für die gesamte Familie auf „0 0“ gesetzt.

Das Software-Paket Genehunter ist im Internet öffentlich zugänglich (Version 2.1 r6, 2006).

8 Kartierung der Marker

8.1 Durchführung der Kartierung als Methode zur Qualitätssicherung

Die Kartierung der verwendeten Marker kann genutzt werden, um mögliche Fehler in den Daten zu entdecken. Aus den genotypisierten Markern wird zu jedem Autosom eine Karte mit den genetischen Abständen in centiMorgan (cM) erstellt.
Chromosomenlängen können als physikalische und genetische Längen angegeben werden. Die physikalischen Längen werden in Basenpaaren (bp) angegeben und entsprechen den beobachtbaren Längen der Chromosomen. In der Kartierung werden die genetischen Längen betrachtet. Das centiMorgan ist ein Maß für die Rekombinationsfrequenz, also für den Austausch genetischer Informationen infolge Trennung benachbarter Genloci. Spricht man bei der Betrachtung zweier Marker davon, dass ihre Loci einen centiMorgan voneinander entfernt liegen, bedeutet dies 1% Wahrscheinlichkeit, dass der Marker am ersten Locus vom zweiten Marker am anderen Locus durch Crossing over während der Meiose getrennt wird. Liegen zwei Marker dicht zusammen, ist diese Wahrscheinlichkeit geringer, als wenn sie weiter voneinander entfernt liegen. Bei Frauen ist die Wahrscheinlichkeit für Rekombinationen größer, somit sind die genetischen Abstände bei ihnen größer und die Karten länger; und die Rekombinationsfrequenzen sind höher als bei Männern (Strachan & Read, 2004).
Auch durch Fehler in den Genotypisierungsdaten werden die Karten länger (siehe Kapitel 4.5).

Die Kartierung kann mit dem Programm *Crimap* erfolgen (Green et al., 1990).
Als Input benötigt *Crimap* eine *.gen* Datei. Diese Datei ähnelt der schon beschriebenen *.ped* Datei.

Zum Stammbaum aus der Abbildung 8.1 zeigt Abbildung 8.2 die entsprechende *.gen* Datei. Die *.gen* Datei enthält Angaben zur Anzahl der enthaltenen Familien und Loci, deren Namen, jeweils der Familienbezeichnung und den Mitgliedern einer Familie. Zu jedem Familienmitglied sind die Person ID, die ID von Mutter und Vater, das Geschlecht und die Allele angegeben. Im Unterschied zur *.ped* Datei bekommt eine weibliche Person beim der Angabe zum Geschlecht in der *.gen* Datei eine „0“ und eine männliche eine „1“ (unbekanntes Geschlecht = „3“).

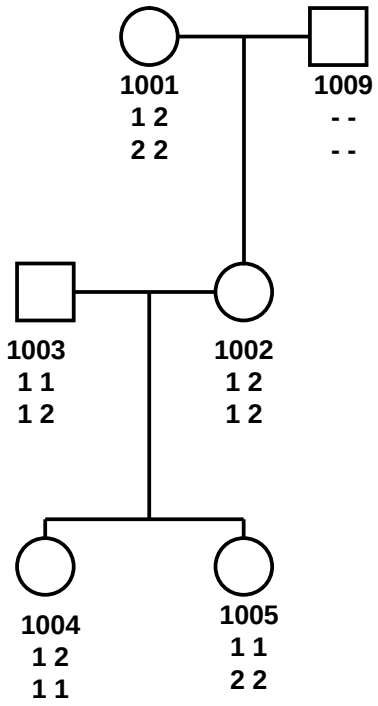

Abbildung 8.1 – Stammbaum zur *.gen* Datei

In diesem Beispiel besteht die *.gen* Datei nur aus einer Familie mit 6 Mitgliedern, die für zwei Marker (LOCA und LOCB) genotypisiert worden sind.

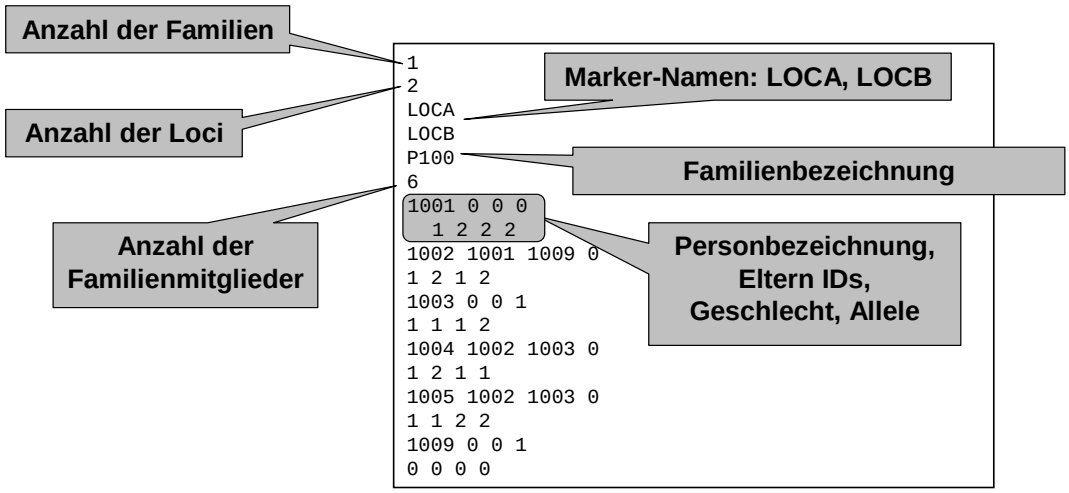

Abbildung 8.2
Beispiel *.gen* Datei

Mit der *prepare* Funktion werden die weiterhin benötigten Dateiformate erstellt: *.loc* Datei, *.ord* Datei, *.par* Datei und *.dat* Datei.
Die *.loc* Datei enthält Informationen zu den Loci: Anzahl der Loci und ihre Namen sowie zwei Spalten zu informativen Meiosen. Dies sind Meiosen, die zu Nachkommen führen und bei der mindestens ein Elternteil mindestens doppelheterozygot ist. Die *.ord* Datei enthält Informationen zu der (An-)Ordnung der Loci.
Die *.par* Datei enthält Werte für die Parameter, die von den verschiedenen *Crimap* Funktionen genutzt werden. Sie ordnet den anderen Dateitypen Namen zu, gibt an, welche Loci schon „eingeordnet" wurden und welche noch eingeordnet werden sollen etc.
Die *.dat* Datei wird durch die *prepare* Option von *Crimap* erstellt und enthält Informationen zu den Haplotypen.
Mit der *build* Funktion kann eine Karte aus den verwendeten Markern erstellt werden. Mit dem Befehl

```
crimap x build > build2x.out,
```

wird schließlich die Ausgabe-Datei erstellt. Anmerkung: Das „x" im Befehl steht für „chrx", also für die Bezeichnung des untersuchten Chromosoms ohne „chr".

Abbildung 8.3 zeigt Teile der *.out* Datei für die Kartierung des Chromosoms 20 in der LRS-Studie. Die Ausgabe enthält eine geschlechtsgemittelte und eine geschlechtsspezifische Karte. Letztere ist in Abbildung 8.3 nicht dargestellt. Für die kartierten Marker sind ihre Position in der Karte, ihre Namen sowie ihre genetischen Abstände vom Startmarker in cM angegeben. Für zwei benachbarte Marker sind jeweils die Rekombinationsfrequenzen zwischen ihnen und ihr Abstand voneinander in cM aufgeführt. Der Marker D20S838 wurde nicht in die Karte eingeordnet. Er steht ganz unten in der Ausgabe und seine möglichen Positionen sind jeweils mit einem „x" gekennzeichnet.

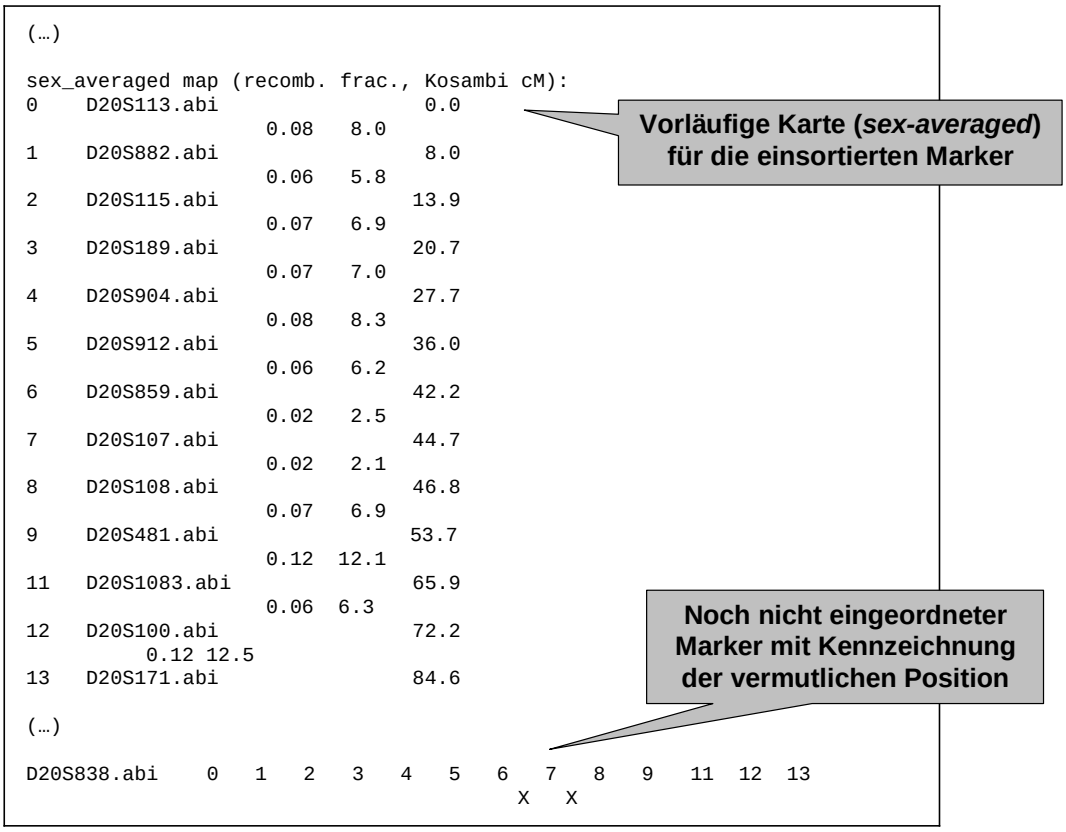

Abbildung 8.3
Ausschnitt aus einer Output-Datei von *Crimap* für das Chromosom 20; dargestellt ist eine geschlechtsgemittelte Karte

Die so erstellten Karten können mit schon vorhandenen Kartierungen verglichen werden, um mögliche Fehler zu finden. Als Referenzkartierungen können beispielsweise die Marshfield Karte oder die Karte von deCode genutzt werden (Broman et al., 1999; Kong et al., 2002).
Drei Aspekte werden vor allem geprüft:

- Stimmen die Chromosomenlängen in etwa überein?
- Lässt sich die Reihenfolge der Marker anhand der Referenzkarten nachvollziehen?
- Liegen alle Marker wirklich auf dem entsprechenden Chromosom?

Die Chromosomenlängen in den selbst erstellten Karten werden immer etwas von den Längen in Referenzkarten abweichen. Die Referenzkarten wurden mit viel mehr Markern und informativeren Meiosen viel feiner kartiert. Dagegen sind die selbst erstellten Karten ungenauer und „roher“. Als Anhalt für Fehler kann der Vergleich der Chromosomenlängen aber dennoch dienen. Wie bereits dargelegt erhöhen Fehler die Kartendistanzen. Würde eine

erstellte Karte von den Referenzkartierungen deutlich abweichen, kann dies durch Fehler bedingt sein.

Ob sich die Reihenfolge der Marker in den Referenzkartierungen nachvollziehen lässt, kann als zweiter Schritt überprüft werden. Liegt der Marker B in der Referenzkartierung zwischen Marker A und Marker C, so sollte er sich auch in der selbst erstellten Kartierung zwischen diesen beiden Markern wieder finden lassen.

Abschließend wird geprüft, ob ein Marker, der in unserer Kartierung verwendet wird, überhaupt auf dem entsprechenden Chromosom liegt. Wenn ein Marker gar nicht auf dem Chromosom liegt, würde man dies daran erkennen, dass der betreffende Marker ganz am Anfang oder am Ende mit einer großen Rekombinationsfrequenz zum nächstgelegenen Marker kartiert wird, und die Kartierung nicht signifikant ist, da keine Kopplung vorliegt.

Die hier beschriebenen Kontrollen werden sämtlich „von Hand“ durchgeführt, dass heißt, die Kartierung wird mit *Crimap* durchgeführt und die Output-Dateien werden für jedes Chromosom geprüft.

An einem Beispiel sollen die beschriebenen Kontrollen verdeutlicht werden. Die Tabelle 8.1 zeigt die Ergebnisse der Kartierung mit *Crimap*, die mit den Markern aus der LRS-Genotypisierung für das Chromosom 20 gemacht wurden.
Für das Chromosom 20 wurden in der LRS Studie 14 Marker genotypisiert. Die Rekombinationsfrequenzen zwischen je zwei benachbarten Markern und der Abstand der Marker vom Startmarker sind in cM angegeben. Als Vergleich sind die Markerpositionen in cM nach der Marshfield Karte und der deCode Karte aufgeführt.

Position in LRS-Kartierung	Marker	Rekombinationsfrequenz in LRS-Kartierung	Abstand vom Startmarker in cM		
			LRS	Marshfield	Decode
0	D20S113		0.0	8.9	7.39
		0.08			
1	D20S882		8.0	15.1	17.8
		0.06			
2	D20S115		13.9	21.2	∅
		0.07			
3	D20S189		20.7	30.6	33.9
		0.07			
4	D20S904		27.7	37.6	41.6
		0.08			
5	D20S912		36.0	46.7	50.9
		0.06			
6	D20S859		42.2	51.4	59.4
		0.02			
7	D20S107		44.7	55.7	61.8
		0.02			
8	D20S108		46.8	57.4	∅
		0.07			
9	D20S481		53.7	62.3	69.4
		0.12			
11	D20S1083		65.9	77.8	∅
		0.06			
12	D20S100		72.2	84.7	88.9
		0.12			
13	D20S171		84.6	95.7	98.6

∅ = nicht enthalten

Tabelle 8.1

Ergebnisse der Kartierung für Chromosom 20

Die angegeben Längen/Abstände differieren, und der Marker D20S113 bekommt in der LRS-Kartierung als Startmarker den Wert 0 cM zugeordnet, während er in der Marshfield Karte und in der deCode Karte einen höheren Wert hat. Dort ist er nicht der Startmarker, und somit beginnt in diesen Kartierungen das Chromosom mit einem anderen Marker. Die Referenzkarten wurden, wie erwähnt, mit viel mehr Markern erstellt. Für Chromosom 20 bestehen die Karte von Marshfield aus 207 Markern und die deCode Karte aus 214 Markern.
Schaut man sich aber die Differenzen zwischen den genetischen Abständen an, wird deutlich, dass die Marker im Vergleich der Karten zueinander ungefähr ähnlich positioniert sind.
Betrachtet man die Rekombinationsfrequenzen, findet man bei den Markern „am Rand“ – am Anfang und am Ende - keine Werte, die überhöht sind und so auf Fehler deuten würden. Desweiteren sind die genetischen Chromosomenlängen ähnlich.
In diesem Beispiel sind die drei eingangs erwähnten Bedingungen erfüllt. Die Chromosomenlänge unserer Kartierung lässt sich mit denen der Referenzkartierungen vergleichen, die Reihenfolge der Marker lässt sich bestätigen und alle Marker liegen auch wirklich auf dem Chromosom.

8.2 Konsequenzen bei Auffälligkeiten in der Kartierung

Es lässt sich nur schwer generalisieren, wie auf Abweichungen bei der Kartierung reagiert werden würde. Treten Abweichungen auf, würde man immer in der Zusammenschau aller Schritte der Qualitätssicherung – der Überprüfung auf Mendel-Fehler und Doppelrekombinanten, der Kartierung und den Tests auf Abweichungen vom Hardy-Weinberg Gleichgewicht und auf Nullallele – entscheiden, ob beispielsweise ein Marker eventuell ganz aus der Analyse genommen würde.

9 Tests unter Nutzung des Hardy-Weinberg Gleichgewicht

Wie in Abschnitt 4.6 beschrieben kann man das Hardy-Weinberg Gleichgewicht nutzen, um nach auffälligen Abweichungen in den Genotypfrequenzen zu suchen. Solche Abweichungen können auf Genotypisierungsfehler hindeuten. So kann das Hardy-Weinberg Gleichgewicht genutzt werden, um erhöhte Raten an Homozygoten zu entdecken, die beispielsweise entstehen können, wenn in der Genotypisierung eines heterozygoten Genotyps nur ein Allel amplifiziert wird.

In diesem Kapitel wird das Prinzip des Hardy-Weinberg Gleichgewichts erklärt – siehe Abschnitt 9.1. Um die Hardy-Weinberg Verteilung nutzen zu können, müssen die Allelfrequenzen und die erwarteten Häufigkeiten an Homozygoten und Heterozygoten bekannt sein. Dies wird in den Abschnitten 9.2 und 9.3 behandelt.

Anschließend sollen die Methoden vorgestellt werden, mit denen die Genotypisierungsdaten auf Abweichungen vom Hardy-Weinberg Gleichgewicht untersucht werden können. Zum einen wird eine χ^2-Statistik verwendet, die im Abschnitt 9.4 beschrieben wird. Mit der χ^2-Statistik wird ein Heterozygotenmangel-Test (engl. *deficiency of heterozygote*, DH-Test) durchgeführt. Zum anderen wird eine Monte-Carlo Permutation verwendet, um erhöhte Raten an Homozygoten in den untersuchten Daten zu entdecken. Diese Permutation ist im Abschnitt 9.5 beschrieben.

Im Abschnitt 9.6 wird ein Test auf das Vorliegen von Nullallelen in den Daten beschrieben.

Eine weitere Methode Abweichungen vom Hardy-Weinberg Gleichgewicht in Genotypisierungsdaten zu erkennen, ist die graphische Darstellung mit den so genannten Vasarely-Diagrammen, auf die in Abschnitt 9.7 eingegangen wird.

9.1 Das Hardy-Weinberg Gleichgewicht

Durch das Gesetz von Hardy und Weinberg wird die Beziehung zwischen Allelfrequenzen und Genotyphäufigkeiten beschrieben. Es besagt, dass diese in aufeinander folgenden Generationen gleich bleiben (Hennig, 2002). Das Hardy-Weinberg Gesetz gilt unter den Bedingungen: zufällige Paarungen, große Populationen, keine geschlechtsspezifischen Unterschiede in den Allelfrequenzen, Allele segregieren nach den Mendel’schen Gesetzen, kein Wirken evolutionärer Kräfte wie Selektion, Drift, Mutation oder ähnliches.

Das Hardy-Weinberg Gleichgewicht soll kurz erläutert werden:

Ein Locus habe zwei Allele. Das Allel A habe die Allelfrequenz p und das Allel B die Allelfrequenz q. Eine willkürlich ausgewählte Person wäre dann homozygot (A/A) mit einer Wahrscheinlichkeit von p^2 (Genotyphäufigkeit P), homozygot (B/B) mit der Wahrscheinlichkeit q^2 (Genotyphäufigkeit Q) oder heterozygot, also (A/B), mit einer Wahrscheinlichkeit von $2pq$ (Genotyphäufigkeit H).

Es gilt $p^2 + 2pq + q^2 = 1$.

Die Allelfrequenzen sind $p + q = 1$.

Die Genotyphäufigkeiten sind $P + H + Q = 1$.

Aus den Frequenzen der Allele (p, q) lassen sich die Frequenzen der möglichen Genotypen (A/A), (B/B) und (A/B) berechnen.

Das Gleichgewicht stellt sich theoretisch schon nach einer Generation ein. Kennt man die Verteilung der Genotypen in der Elterngeneration, kann man auch die Verteilung der nächsten Generation, der Kinder bestimmen. Ebenso lässt sich bei zwei Allelen bei bekannter Häufigkeit eines Allels / einer (homozygoten) Genotyphäufigkeit die Häufigkeit des anderes Allels / der anderen Genotyphäufigkeit berechnen.

Ein Nutzen des Hardy-Weinberg Gleichgewicht besteht darin, Populationen auf Abweichungen von ihm zu testen. Zeigen sich Abweichungen, dann ist wenigstens eine der oben gemachten Annahmen falsch. Erhöhte Raten an Homozygoten können auf Nullallelen oder anderen Genotypisierungsfehlern beruhen. Findet man in den Genotypisierungsdaten erhöhte Raten an Homozygoten, kann auf Fehler geschlossen werden.

Abweichungen vom Gleichgewicht können aber auch aus verschiedenen anderen Gründen resultieren. Ein Heterozygotenmangel kann durch den Wahlund Effekt begründet sein. Dieser Wahlund Effekt besagt, dass in einer Sub-Populationen mehr Homozygote auftreten, als in einer großen Population, die aus mehrere Subpopulationen zusammengesetzt ist (Ridley, 2003). Nah verwandte Eltern (Konsanguinität) können ebenso zu einem Heterozygotenmangel führen wie assortative Paarung, also wenn beispielsweise große Menschen eher große Menschen heiraten oder Taubstumme eher Taubstumme (Seyffert, 1998). Ein weiterer wichtiger Grund für Abweichungen vom Hardy-Weinberg Gleichgewicht kann sein dass der Genotyp das Überleben beeinflusst, also zum Beispiel Heterozygote überleben eher als Homozygote.

9.2 Schätzen der Allelfrequenzen

Um das Hardy-Weinberg Gleichgewicht nutzen zu können, müssen die Frequenzen der Allele bekannt sein. Es existieren geschätzte Allelfrequenzen für viele Marker, die aus zufälligen Stichproben von nicht verwandten Personen stammen. Diese Allelfrequenzen kann man etwa den öffentlich zugänglichen Genotypisierungskarten von Genethon (Dib et al., 1996), CHLC (1999) oder Marshfield (Broman et al., 1998) entnehmen. Für eine grobe Nährung sind diese vorhandenen Werte nutzbar.
Da sich jedoch die Allelfrequenzen bei verschiedenen Populationen deutlich unterscheiden können, ist es besser, die Allelfrequenzen aus den tatsächlich untersuchten Personen zu ermitteln (Terwilliger & Ott, 1994).

Die Allelfrequenzen werden durch Auszählen der Allele bestimmt. Die Allelfrequenz $\hat{p}_A$ für ein Allel A bei einer untersuchten Gruppe von Personen ergibt aus der Anzahl der Personen, die ein Allele A haben (falls sie homozygot (A/A) sind, wird doppelt gezählt), geteilt durch die doppelten Anzahl an Personen, die für den betreffenden Marker genotypisiert sind. Die allgemeine Formel hierzu ist:

$$\hat{p}_i = \frac{2\#(A_iA_i) + \sum_{i \neq j} \#(A_jA_i)}{2N}$$

Das Programm *Pedstats* (Version 0.6.4), auf das weiter unten näher eingegangen wird, arbeitet mit dieser Formel. A_i steht hierbei für die jeweilige Anzahl der verschiedenen Allele des Markers. n ist die Anzahl der Personen, die für den untersuchten Marker genotypisiert sind. Für einen Marker mit den Allelen A, B und C, müsste die Formel zur Bestimmung der Allelfrequenz $\hat{p}_A$ lauten:

$$\hat{p}_A = \frac{2\#(AA) + \#(AB) + \#(AC)}{2n}$$

Eine weitere Möglichkeit zur Bestimmung der Allelfrequenzen bietet das Programm *Downfreq* (Version 1.1, 1995), auf das hier nicht näher eingegangen werden soll.

9.3 Erwartete Häufigkeiten unter Hardy-Weinberg Gleichgewicht

Die erwarteten Häufigkeiten an Heterozygoten und Homozygoten werden mit den bekannten Frequenzen ermittelt. So ist die Anzahl der erwarteten Homozygoten $e_{(hom)}$ das Produkt aus

der Gesamtzahl der untersuchten Personen n und der erwarteten Homozygotenfrequenz $p_{(hom\ exp)}$:

$$e_{(hom)} = p_{(hom\ exp)} * n$$

Die erwartete Homozygotenfrequenz ergibt sich nach dem Hardy-Weinberg Gesetz

$$p^2 + 2pq + q^2 = 1$$

aus dem Quadraten der Allelfrequenzen p und q, deren Bestimmung im vorangegangenen Abschnitt erläutert wurde.

Die erwartete Anzahl an Heterozygoten wird nach der Formel

$$e_{(het)} = n - e_{(hom)}$$

bestimmt.

9.4 χ^2-Test zum Hardy-Weinberg Gleichgewicht

Um einen Heterozygotenmangel-Test durchzuführen, soll eine Chi-Quadrat Statistik verwendet werden.

Der χ^2-Test kommt oft dann zur Anwendung, wenn man tatsächlich vorliegende Häufigkeiten mit den Häufigkeiten vergleichen möchte, die man erwartet. Es handelt sich um einen approximativen Test. Das heißt, die exakte Überschreitungswahrscheinlichkeit p wird mit wachsendem Stichprobenumfang genauer geschätzt.

Für kleinere Datensätze müssen exakte Testverfahren angewendet werden, für hinreichend große ist der χ^2-Test anwendbar. Es wird die Nullhypothese (keine Abweichung) gegen die Alternativhypothese (Abweichung) geprüft.

Mit dem χ^2-Test soll nun geprüft werden, ob es einen Homozygotenexzess gibt, also eine Abweichung der Verteilung von Homozygoten und Heterozygten auftritt. Die erwarteten Zahlen ergeben sich nach dem Hardy-Weinberg Gesetz (bei bekannten Allelfrequenzen).

Zur Demonstration des χ^2-Tests soll der Marker D1S468 dienen, der in der LRS-Genotypisierung untersucht wurde. Tabelle 9.1 zeigt die beobachtete und erwartete Verteilung von Heterozygoten und Homozygoten für diesen Marker.

	Heterozygote	**Homozygote**	**n**
Beobachtet	710	297	1007
Erwartet	733	274	1007

Tabelle 9.1

Verteilung von Heterozygoten und Homozygoten für den Marker D1S468

n sind hier die für einen Marker genotypisierte Personen, die sich auf k Kategorien – homozygot oder heterozygot – verteilen.

Die Nullhypothese H_0 ist, dass die Zahlen der tatsächlich beobachteten Homozygoten und Heterozygoten übereinstimmen mit den Zahlen der nach der Hardy-Weinberg Verteilung erwarteten Homozygoten und Heterozygoten. Die Alternativhypothese H_1 lautet: Es liegt eine Abweichung zwischen beobachteten und erwarteten Zahlen vor - und zwar aufgrund von Nullallelen am ehesten eine Abweichung als Homozygotenexzess / Heterozygotenmangel. Die beiden Hypothesen H_0 und H_1 sind zweiseitig bzw. werden zweiseitig geprüft, d.h. es werden Abweichungen in beide Richtungen („zuviel“ und „zuwenig“) analysiert.

Die Gesamtzahl n und die erwartete Homozygotenfrequenz $p_{(hom\ exp)}$ werden als bekannt vorausgesetzt. $p_{(hom\ exp)}$ kann auch aus den Allelfrequenzen berechnet werden. Die Zahl der erwarteten Homozygoten $e_{(hom)}$ ergibt sich aus der Gesamtzahl n und der Homozygotenfrequenz $p_{(hom\ exp)}$ zu

$$e_{(hom)} = p_{(hom\ exp)} * n\ .$$

Die tatsächlich beobachteten Homozygoten $b_{(hom)}$ werden durch Auszählung ermittelt.

Die Anzahl der erwarteten Heterozygoten $e_{(het)}$ wird ermittelt aus

$$e_{(het)} = n - e_{(hom)}$$

Die tatsächlich beobachteten Heterozygoten $b_{(het)}$ ergibt sich aus

$$b_{(het)} = n - b_{(hom)}$$

Der χ^2-Test sieht dann wie folgt aus

$$\chi^2_{FG=1} = \frac{(b_{(hom)} - e_{(hom)})^2}{e_{(hom)}} + \frac{(b_{(het)} - e_{(het)})^2}{e_{(het)}}$$

Die Anzahl der Freiheitsgrade (FG) bestimmt sich aus der Anzahl der frei variierbaren Summanden. In dieser Formel ist dies ein Summand (FG = 1).

Das Ergebnis χ^2 ist der empirische Wert, der mit dem entsprechenden Quantil aus der χ^2-Verteilung verglichen wird. Die Nullhypothese ist zu verwerfen wenn der kritische χ^2-Wert kleiner als der empirische χ^2-Wert ist $(\chi^2_{crit} < \chi^2_{emp})$.

Für den Marker D1S468 soll für n = 1007 Individuen die Anzahl der Homozygoten ermittelt werden, um zu sehen, ob diese von der erwarteten Anzahl abweicht. Die erwartete Homozygotenfrequenz $p_{(hom\ exp)}$ sei bekannt, 0.27. Es wird ein Signifikanzniveau von $\alpha = 0.05$ festgesetzt.

Die Anzahl der erwarteten Homozygoten ergibt sich dann zu $e_{(hom)}$ = 274. Die erwartete Anzahl an Heterozygoten ist $e_{(het)}$ = 733. Die tatsächlichen Homozygoten und Heterozygoten werden bestimmt. Es wird in die Formel χ^2 eingesetzt:

$$\chi^2_{FG=1} = \frac{(297 - 274)^2}{274} + \frac{(710 - 733)^2}{733}$$

$$\chi^2_{1FG} = 2.43$$

Für einen Freiheitsgrad FG = 1 und das Signifikanzniveau $\alpha = 0.05$ gilt ein Grenzwert von 3.84. Der erhaltene χ^2-Wert liegt darunter und erlaubt in diesem Beispiel also eine Beibehaltung der Nullhypothese, dass es keinen Homozygotenexzess gibt.

Zu diesem Ergebnis kann ein empirischer p-Wert berechnet, der sich zu 0.11 ergibt.

Der χ^2-Test wird mit dem Programm *Pedstats* realisiert. Es wird für alle Marker überprüft, ob sich bei ihnen signifikante Abweichungen der Allelfrequenzen vom HWE ergeben.

Pedstats benötigt als Inputdateien eine *.ped* und eine *.dat* Datei.

Pedstats führt entweder einen approximativen χ^2-Test oder einen exakten Test durch. Dies hängt von der Anzahl der Allele eines Markers ab. Um zu kleine Zellbesetzungen zu vermeiden, wird ein *Pooling* Algorithmus durchlaufen. Hierbei werden die Allele gruppiert und Allele mit sehr kleinen Allelfrequenzen zu einer Gruppe zusammengefasst. Das Programm durchläuft folgende Schritte:

- Die Allelfrequenzen $\hat{p}_i$ werden durch Allelauszählung bestimmt (Abschnitt 9.2.)
- Alle Allele werden entweder als „selten“

  $$\hat{p}_i < \sqrt{\frac{3}{n}}$$

  oder als „häufig“

$$\hat{p}_i \geq \sqrt{\frac{3}{n}}$$

klassifiziert.

- Wenn alle Allele selten sind, wird das Allel mit der höchsten Frequenz aus dem Pool der seltenen Allele herausgenommen und in einem eigenen Pool platziert, damit kein Test mit nur einem Pool durchgeführt wird.
- Alle Allele im Pool der „seltenen“ ($A_i \in R$) werden als eine Gruppe von Allelen (quasi als ein Allel) aufgefasst, deren Frequenz $\hat{p}_R$ erneut bestimmt wird.

Nach diesem *Pooling* wird entweder ein χ^2-Test für jeden Marker mit mehr als zwei Allelen durchgeführt oder ein exakter Test, wenn ein Marker nur zwei Allele hat.
Pedstats benutzt für den χ^2-Test folgende Formel (Verallgemeinerung der Formel für χ^2 von oben) für die Teststatistik:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{i} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

wobei m die Zahl der Allele des Markers bezeichnet, A_i ist das i^{te} Allel. E_{ij} bezeichnet die erwartete Allelfrequenz des Genotypen A_iA_j und O_{ij} die beobachtete Häufigkeit von A_iA_j.
Unter der Nullhypothese, nach der die Daten im Hardy-Weinberg Gleichgewicht sind, gilt

$$E_{ij} = \begin{cases} 2n\hat{p}_i\hat{p}_j, \text{wenn } i \neq j; \\ n\hat{p}_i^{\,2}, \text{wenn } i = j. \end{cases},$$

wobei n die Anzahl der Personen ist, die für den Locus genotypisiert sind und $\widehat{p_i}$ die geschätzte Allelfrequenz von A_i ist.
Für große N ist die Teststatistik χ^2-verteilt mit $m(m-1)/2$ Freiheitsgraden für einen Marker mit m Allelen. Für kleine Fallzahlen bzw. erwartete Häufigkeiten E_{ij} ist die Approximation durch die χ^2-Verteilung ungenau. In diesem Fall, würde *Pedstats* den Test zwar durchführen, aber die Ergebnisse dementsprechend kennzeichnen.

Mit dem Programm *Pedstats* werden für jedes Chromosomen alle Marker bei den *Foundern* auf ihre Kompatibilität mit dem Hardy-Weinberg Gleichgewicht überprüft.

Als Beispiel soll die Prüfung mit *Pedstats* für das Chromosom 1 aus der LRS-Genotypisierung dienen.

Die Befehlszeile sieht wie folgt aus:

```
pedstats -p chr1_ph_nwl.pre -d chr1_ph_nwl.dat --hardyweinberg --checkfounders
```

Die Abbildung 9.1 zeigt die Bildschirmausgabe für diese Kontrolle.

```
HARDY-WEINBERG CHECK AMONG FOUNDERS
===================================
                  N_HOM  N_HET  E_HET      N_ALLELES ALLELES P-VALUE
D1S2713             120    372    342   4, 11 pooled    1-14  0.0110 A
D1S2688             192    299    303   3,  2 pooled     1-4  0.0120 A
```

Abbildung 9.1

Bildschirmausgabe von *Pedstats*

Es werden die zwei Marker aufgeführt, die in der Kontrolle auffällige p-Werte ergaben. Abgebildet ist jeweils der Marker-Name, die Zahl der tatsächlichen Homo- und Heterozygoten, die Zahl der erwarteten Heterozygoten, die Zahl der Marker im Test (beim oberen Marker 4, weil 11 Marker zusammengefasst sind), alle Allele für den Marker (oberer Marker: 14) und der p-Wert für den durchgeführten χ^2-Test. Die Buchstaben („A“) bedeuten, dass asymptomatische χ^2-Tests durchgeführt wurden und nicht exakte Tests, die mit einem („E“) gekennzeichnet wären.

In diesem Beispiel wurden die zwei Marker vom Programm ausgegeben, weil ihre p-Werte kleiner als 0.05 waren.

Pedstats bietet auch die Möglichkeit, die Ergebnisse der Hardy-Weinberg Analyse graphisch nachzuvollziehen.

Erweitert man die Befehlszeile von oben um *--pdf*

```
pedstats -p chr1_ph_nwl.pre -d chr1_ph_nwl.dat --hardyweinberg --checkfounders --pdf,
```

wird eine *.pdf* Datei erstellt, die neben Informationen zu Familienstrukturen, zum Betroffenheitsstatus, zu Loci und Allelfrequenzen auch die Auswertungen für die Hardy-Weinberg Analyse graphisch darstellt. Die Abbildung 9.2 zeigt diese graphische Darstellung für den Marker D1S2688.

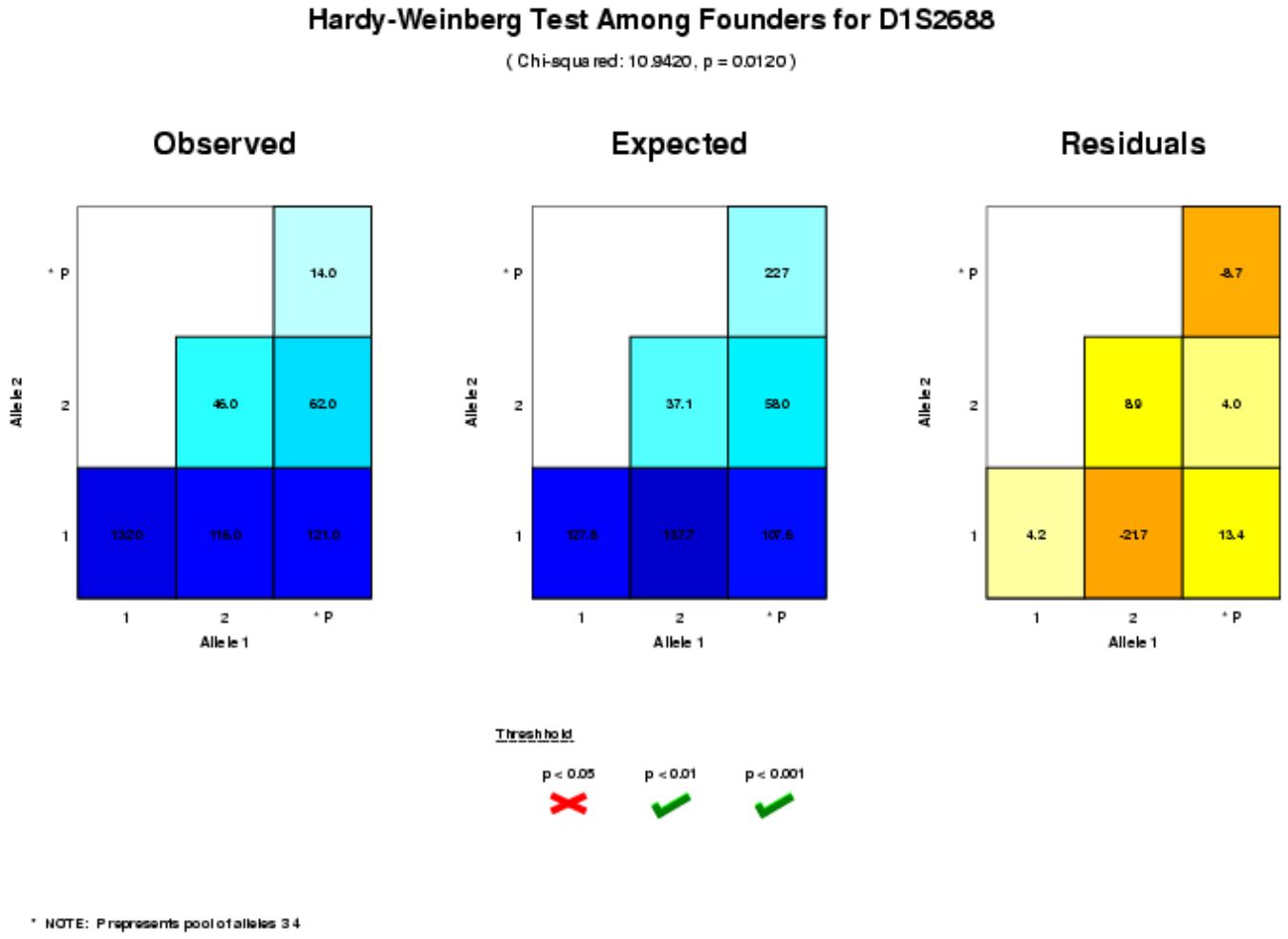

Abbildung 9.2

Graphische Darstellung der Hardy-Weinberg Analyse für den Marker D1S2688

Unter dem Titel sind der χ^2 Wert und der zugehörige p-Wert abgebildet. Der p-Wert deutet auf Abweichung vom Hardy-Weinberg Gleichgewicht. Die linke Tafel zeigt die beobachtete Allelverteilung ("Observed"). Auf der waagerechten Achse ist das erste Allel, auf der senkrechten Achse das zweite Allel aufgetragen. Die Allele 3 und 4 wurden als seltene Allele zu einem Allel zusammengefasst (p).

Die Werte in den Zellen repräsentieren die Anzahl der beobachteten Genotypen. Der Genotyp (2/2) wurde zum Beispiel 46mal beobachtet. Zellen von Genotypen, die häufig erwartet werden, sind dunkelblau, Zellen von selteneren Genotypen sind heller dargestellt.
Die mittlere Tafel („Expected“) gibt die nach der Hardy-Weinberg Verteilung erwartete Allelverteilung wieder. Die Werte in den Zellen sind die erwarteten Genotyphäufigkeiten. Die Färbung richtet sich wie bei der Tafel der Beobachteten zuvor nach der erwarteten Zahl an Genotypen.
Schaut man sich die beiden Tafeln im Vergleich an, so sieht man, dass die Anzahl der beobachteten homozygoten Genotypen (1/1) der Anzahl der erwarteten relativ nahe kommt. Die Anzahl der beobachteten Heterozygoten (1/2) weicht dagegen schon deutlich von der Anzahl der erwarteten Heterozygoten ab.
Die rechte Tafel (Residuals) zeigt jeweils die Werte für die Differenz zwischen beobachteten und erwarteten Genotypen an. So ist die Differenz z.B. beim homozygoten Genotyp (1/1) 4.2, beim heterozygoten Genotyp (1/2) 21.7.
Die Färbung der Zellen in der rechten Tafel richtet sich danach, ob es signifikante Abweichungen vom Hardy-Weinberg Gleichgewicht gibt oder nicht. Nach der Formel

$$Residuals_{ij} = \frac{(Observed_{ij} - Expected_{ij})}{\sqrt{Expected_{ij}}}$$

geht der Wert gegen 0, wenn es wenig oder keine Abweichungen gibt.
Die Anzahl der Homozygoten (1/1) weicht kaum ab, weshalb hier die Färbung hell ist, die Anzahl der heterozygoten (1/2) weicht dagegen stärker ab, weshalb die Färbung der Zelle kräftiger ist.

Das Programm Pedstats ist im Internet kostenfrei verfügbar.

9.5 Monte-Carlo Permutation zum Hardy-Weinberg Gleichgewicht

Um Abweichungen von der Hardy-Weinberg Verteilung zu entdecken, kann auch eine Monte-Carlo Permutation durchgeführt werden. Unter einer Permutation (lat. *permutare* = (ver)tauschen) versteht man die Veränderung der Anordnung einer Menge durch Vertauschen ihrer Elemente. In der Monte-Carlo Permutation werden von allen Genotypen die zwei Allele als einzeln stehend betrachtet. Sie werden zu neuen Allelpaaren zusammengesetzt. Die so permutierten neuen Allelpaare werden wiederum untersucht.

Die zu untersuchende n Genotypen liegen mit einer beobachteten Häufigkeit von Homozygoten Hom_O vor. Man kann die n Genotypen als eine Menge von 2n Gameten ansehen, die zufällig zusammengesetzt sind.

A_1	$\#_{11}$			
A_2	$\#_{12}$	$\#_{22}$		
…	…	…	…	
A_m	$\#_{m1}$	$\#_{m2}$	…	$\#_{mm}$
	A_1	A_2	…	A_m

Unter der Annahme des Hardy-Weinberg Gleichgewichtes können die n Genotypen mit den Gameten $\#_i\ A_i$ als n zufällig verteilte Verbindungen von zwei Gameten angesehen werden.

Die 2n einzelnen Gameten können so betrachtet werden, wie sie „vor“ der zufälligen Verteilung vorlagen:

$\vert\leftarrow \#_1 \rightarrow\vert$			$\vert\leftarrow \#_2 \rightarrow\vert$						$\vert\leftarrow \#_m \rightarrow\vert$			
A_1	A_1	…	A_1	A_2	A_2	…	A_2	…	…	…	A_m	A_m	…	A_m

Die Monte-Carlo Methode setzt an diesem Punkt an. Die untersuchte Stichprobe von n Genotypen, wird als eine Menge von 2n einzelner Gameten angesehen. Diese Menge wird gemischt und zufällig wieder zu n „neuen Genotypen“ zusammengesetzt - permutiert. Die neu zusammengesetzte Gruppe von Genotypen kann mit der beobachteten Gruppe verglichen werden. In der neuen Gruppe werden wiederum die Anzahl der homozygoten Personen Hom_P bestimmt und mit der vorher beobachteten Anzahl Hom_O verglichen.

Der Algorithmus dazu ist folgender:

- Berechnung der Anzahl der homozygoten Personen Hom_O.
- Zähler C = 0 setzen.
- Für N Permutationen wiederholen der folgenden Schritte:
    - permutieren der Allele und Neuverteilung zu neuen Allelpaaren,
    - Berechnung der Anzahl der homozygoten Personen im permutierten Datensatz Hom_P,
    - Vergleich: $Hom_O \leq Hom_P$, wenn dies vorliegt, Addition von 1 zu.
- Der Monte-Carlo p-Wert ist C/N.

Sind die untersuchten Genotypen kompatibel zum Hardy-Weinberg Gesetz, so sollten Hom_O und Hom_P ähnlich sein. Liegen in den untersuchten Genotypen mehr Homozygote vor als nach der Hardy-Weinberg Verteilung erwartet, wird der Monte-Carlo p-Wert klein werden, d.h. er wird Werte nahe 0 annehmen. Sind die Genotypen im Hardy-Weinberg Gleichgewicht oder liegen mehr Heterozygote (also weniger Homozygote Hom_O) vor, wird der Monte-Carlo p-Wert einen Wert zwischen 0.5 und 1 annehmen.
Eine bestimmte Anzahl von Mutationen sollte durchgeführt werden, um eine vorher bestimmte Genauigkeit des geschätzten p-Wertes sicherzustellen. Unter Benutzung des zentralen Grenzwertsatzes, kann eine asymptotische Normalverteilung des geschätzten p-Wertes angenommen werden. Die Zahl der nötigen Permutationen N um einen p-Wert zu schätzen, der kleiner als π mit einer Genauigkeit von e in einem Konfidenzintervall von $1-\alpha$ ist, ergibt die Formel

$$N = \left(\frac{z_{1-\alpha/2}(\pi(1-\pi))^{1/2}}{e} \right)^2 .$$

$\pi(1-\pi)$ hat sein Maximum bei $\pi = \frac{1}{2}$, so dass sich für praktische Zwecke die Formel reduzieren lässt zu

$$N = \left(\frac{z_{1-\alpha/2}}{2e} \right)^2 ,$$

wenn keine a priori Wahrscheinlichkeit über den p-Wert bekannt ist.
Wenn eine p-Wert mit einer Präzision von $e = 0.01$ mit 99% Konfidenz geschätzt werden soll, und mit dem $1-\alpha/2 = 0.995$ Quantil der Standardnormalverteilung von 2.576, sind ungefähr

$$N = \left(\frac{2.576}{2*0.01} \right)^2 \approx 16589.44 \approx 17000$$

Simulationen nötig (Ziegler und König, 2006).

In der Untersuchung der Daten aus der LRS-Studie werden jeweils 20000 Permutationen durchgeführt.

Ein Beispiel soll den Test mit der Monte-Carlo Permutation verdeutlichen.

Ausgehend vom Hardy-Weinberg Gleichgewicht sollen die biallelen Marker von 100 fiktiven Genotypen betrachtet werden. Die Allele a und b haben die Frequenz p_a = 0.75 und p_b = 0.25. Nach dem Hardy-Weinberg Gleichgewicht ist ein bialleler Marker wie folgt verteilt

$$a^2 + 2ab + b^2 .$$

Es ergeben sich a² = 0.5625 (im Beispiel 56 von 100 Personen), b² = 0.0625 (6 von 100 Personen) und 2ab = 0.375 (38 von 100 Personen).

Im Gleichgewicht sind also 62 Personen homozygot und 38 Personen heterozygot.

Die Monte-Carlo Permutation ergibt bei N = 20000 einen p-Wert von p = 0.6234

Der Plot der 20000 permutierten Homozygoten ist in Abbildung 9.3 dargestellt.

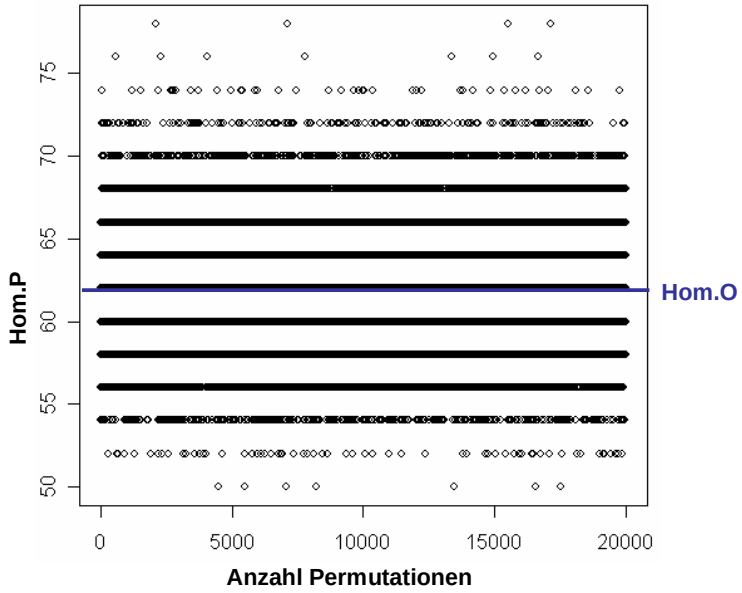

Abbildung 9.3

Plot einer Permutation bei Hardy-Weinberg Gleichgewicht

Die Anzahl der beobachteten Homozygoten Hom_O ist durch die blaue Linie wiedergegeben. Jede Permutation ist durch jeweils einen Punkt dargestellt. Auf der x-Achse sind die

Permutationen von 1 bis 20000 aufgetragen, auf der y-Achse die Anzahl an permutierten Homozygoten Hom_P.

Man erkennt, dass die meisten permutierten Werte um oder auf dem Wert für Hom_O liegen.

Betrachtet man nun die Genotypen von 100 fiktiven Personen die sich nicht der Hardy-Weinberg Verteilung entsprechen befinden, zum Beispiel 80 Homozygote und 20 Heterozygote, ergibt sich ein Monte-Carlo p-Wert von p = 0.0013. Der entsprechende Plot ist in Abbildung 9.4 dargestellt.

Die Anzahl der beobachteten Homozygoten Hom_O ist wiederum durch den blauen Strich gekennzeichnet. Diesmal liegen aber der größte Teil der permutierten Werte unter dem Wert von Hom_O.

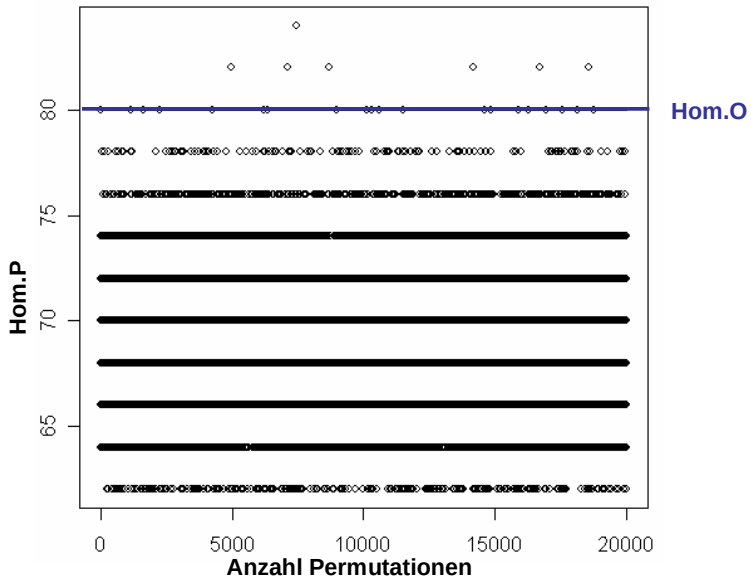

Abbildung 9.4
Beispiel für einen Plot bei Homozygotenexzess

Bei Genotypen, unter denen sich vermehrt Heterozygote befinden, und damit weniger Homozygote als erwartet, ergeben sich p-Wert gegen 1. Bei diesem letzten Beispiel liegen 50 Homozygote und 50 Heterozygote vor, es ergibt sich ein Monte-Carlo p-Wert von p=0.95065. Die Abbildung 9.5 zeigt den Plot für dieses Beispiel.

In der Abbildung ist zu erkennen, dass der größte Teil der permutierten Werte diesmal oberhalb des Wertes für Hom_O liegt. Bei den meisten Permutationen liegen mehr Homozygote vor als die erwarteten 50, was auf einen Heterozygotenexzess hindeutet.

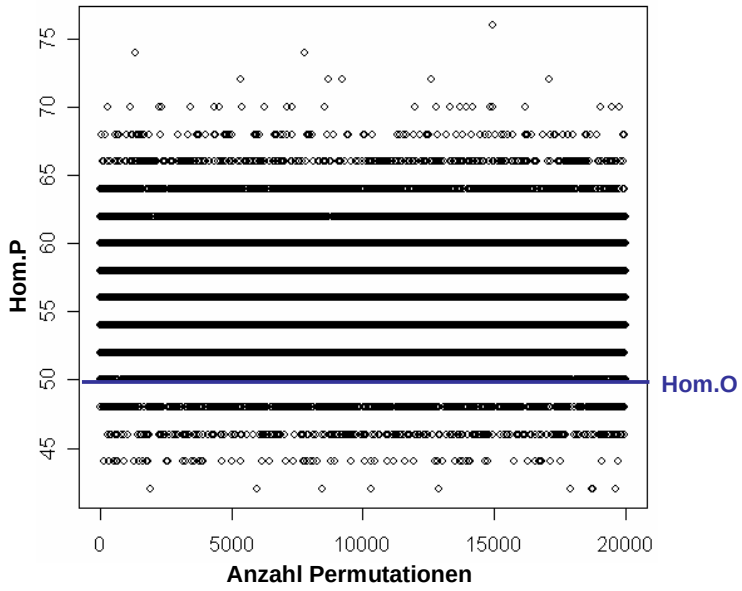

Abbildung 9.5
Beispiel für einen Plot bei Heterozygotenexzess

Die Monte-Carlo Permutation ist als Algorithmus in der Programmiersprache *R* implementiert.
Als Input kann eine *.ped* Datei im bekannten Format benutzt werden. Die Ausgabe umfasst den Monte-Carlo p-Wert, den Wert der beobachteten Homozygoten Hom_O sowie einen Plot der permutierten Werte wie in den Abbildungen 9.3, 9.4 und 9.5 dargestellt.
Wie zuvor beim χ^2-Test wird die Monte-Carlo Permutation nur für die *Founder* durchgeführt.

9.6 Test auf Nullallele

Genotypisierungsfehler, die auf Nullallelen basieren, können zu Abweichungen vom Hardy-Weinberg Gleichgewicht führen. Zum Test auf Nullallele wird das Programm *NullAlleleCheck* verwendet (Version 0.98, 2004). Der Test wird für jedes Chromosom für alle Marker durchgeführt.
Als Input benötigt *NullAlleleCheck* eine *.ped* und eine *.dat* Datei. Die *.dat* Datei muss Allelfrequenzen enthalten, weil sie für die Berechnung gebraucht werden. Diese Allelfrequenzen können aus den genotypisierten Familien geschätzt werden. Für jeden Marker werden die folgenden Analysen durchgeführt, die schon aus Abschnitt 9.4 bekannt sind:

- χ^2 Anpassungstest für Nullallele

$$\chi^2_{FG=1} = \frac{(b_{(hom)} - e_{(hom)})^2}{e_{(hom)}} + \frac{(b_{(het)} - e_{(het)})^2}{e_{(het)}}$$

- Berechnung des p-Wert

Eine weitere Möglichkeit der Überprüfung stellt der ML-(*Maximum likelihood*-)Schätzer für die Frequenz der Nullallele $\hat{p}_x$ nach Fisher dar (Fisher et al., 2001).

- *NullAlleleCheck* bestimmt diesen Wert

$$\hat{p}_x = \frac{(e_{(het)} - b_{(het)})}{(e_{(het)} + b_{(het)})}$$

Für den ML-Schätzer gilt ein Wert unter 0.02 als normal, ein Wert großer 0.02 wäre signifikant für das Vorliegen von Nullallelen (Fisher et al., 2001; Ziegler mündlich).

Als Beispiel soll der Marker D1S468 aus dem Beispiel von oben dienen.
Der Wert für χ^2 wurde bereits errechnet $\chi^2_{FG=1} = 2.43$.
Für den ML-Schätzer ergibt sich ein Wert von:

$$\hat{p}_x = \frac{(733-710)}{(733+710)} = \frac{23}{1443} = 0.0159$$

Die Bildschirm-Ausgabe ist in der Abbildung 9.6 wiedergegeben. Sie enthält jeweils den Marker-Namen Marker, die Anzahl der Genotypen ohne Nullen N, die beobachteten und erwarteten Homozygoten Hom(Obs) und Hom(Exp), sowie χ^2-Wert, p-Wert und $\hat{p}_x$-Wert.

```
alex@bioinf:~> nullac -p chr1_ph_nwl.pre -d chr1_ph_nwl.dat
    Marker  N    Hom(Obs) Hom(Exp)   X^2    p-Value    pHatX
    D1S468  1007 297      273.914  2.67282  0.102075   0.0159979
```

Abbildung 9.6
Ausgabe von *NullAlleleCheck 0.98*

Es wurde schon festgestellt (siehe Kapitel 9.4), dass es an diesem Marker keinen Homozygotenexzess gibt. Auch der ML-Schätzer für die Nullallelfrequenzen bleibt unter

0.02. Bei diesem Marker würde man also annehmen, dass keine Störung durch Nullallele vorliegt.

9.7 Vasarely-Abbildungen

In den vorangegangenen Abschnitten wurde beschrieben, wie man Nullallele und Abweichungen vom Hardy-Weinberg Gleichgewicht entdecken kann. Die bisher beschriebenen Methoden sind gut geeignet für den automatisierten Test großer Datenmengen, dass heißt, vieler Marker mit vielen Allelen.

Eine weitere Methode, Abweichungen von der Hardy-Weinberg Verteilung zu untersuchen, ist die visuelle Darstellung anhand von Vasarely-Tafeln. Manaster wandte erstmals 1999 Vasarely-Tafel nach dem Vorbild des Künstlers Victor Vasarely (1908-1997) zur Darstellung von Abweichungen vom Hardy-Weinberg Gleichgewicht an (Manaster, Nanthakumar et al., 1999).

Die Methode nutzt die graphische Darstellung der Allelfrequenzen.

Die Abbildung 9.7 zeigt die Vasarely Darstellung für den Marker D7S2429 aus der Dyslexie Studie.

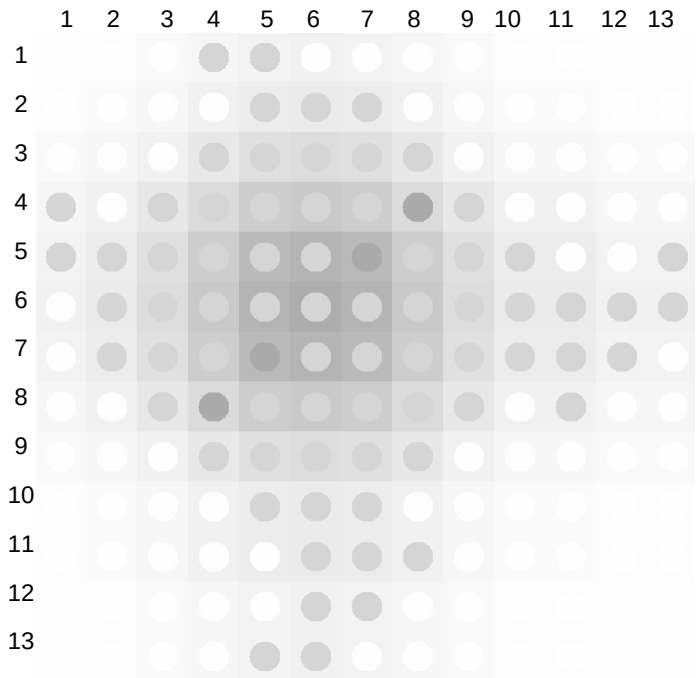

Abbildung 9.7

Vasarely-Tafel für einen Marker aus der LRS-Studie

Der Marker besitzt 13 Allele. Die waagerechten Reihen stellen das erste Allel dar, die senkrechten Spalten das zweite.

Die Abbildung zeigt für jede mögliche Allelkombination die gemessenen und erwarteten Allelfrequenzen. Die gemessenen Allelfrequenzen sind durch Kreise dargestellt, die in der Mitte von Quadraten liegen. Diese Quadrate stellen die erwarteten Allelfrequenzen dar. Besteht kein signifikanter Unterschied zwischen erwarteten und gemessenen Allelfrequenzen – also liegen keine Nullallele oder andere störende Effekte vor – haben Kreise und Quadrate dieselbe oder ähnliche Färbung und werden nicht als unterschiedlich wahrgenommen.

Liegen nun Abweichungen, wie z.B. Nullallele vor, unterscheiden sich beobachtete und erwartete Allelfrequenzen und somit auch die Farben/Graustufen. Die Kreise heben sich deutlich von ihrem Hintergrund ab.

Beim dargestellten Marker fallen so zum Beispiel die Allelkombinationen (1/4), (1/5) oder (4/8) auf. Die gemessenen Frequenzen für diese Kombinationen heben sich deutlich von den erwarteten Frequenzen für diese Kombinationen ab. Leichte Abweichungen im Sinne eines Überschusses ein Homozygoten weisen das Allel 4 und das Allel 8 auf.

Ein Homozygotenmangel ist beispielsweise für das Allel 6 zu beobachten. Der helle Kreis vor dem dunkleren Quadrat zeigt an, dass es weniger Homozygote (6/6) gibt als erwartet.

Die Abbildung 9.8 zeigt einen Marker mit 13 Allelen, bei dem Nullallele vorliegen.

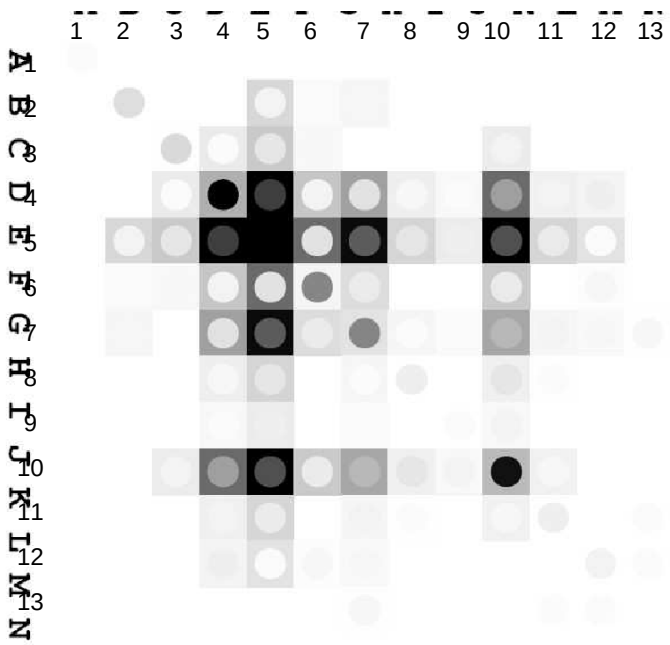

Abbildung 9.8

Vasarely-Tafel für einen Marker mit Nullallelen – angelehnt an Manaster (1999)

Es ist zu erkennen, dass die Frequenzen der gemessenen Homozygoten für die Allele 2, 3, 4, 6, 7, 8 und 10 sich deutlich von den erwarteten Frequenzen unterscheiden. Der sichtbare Unterschied in den Farben/Graustufen weist auf Abweichungen vom Hardy-Weinberg Gleichgewicht hin - beispielsweise aufgrund erhöhter Homozygotie.

Für die Kontrolle von Genotypisierungsdaten mit Hilfe der Vasarely Tafeln existiert von Manaster ein Windows Programm. Die benötigten Dateiformate lassen sich mit einem *R*-Tool aus *.ped* Dateien generieren.
Die Ausgabe erfolgt als Graphik unter der *Windows* Oberfläche.
Die Vasarely Methode kommt für die Untersuchungen im Rahmen der LRS-Studie nicht zur Anwendung, weil sie für große Datensätze unpraktikabel ist. Die Interpretation der Vasarely Tafeln ist darüber hinaus jeweils Betrachter abhängig.

9.8 Konsequenzen aus den Tests zum Hardy-Weinberg Gleichgewicht

Die Genotypisierungsdaten werden mit den oben beschriebenen Kontrollen auf ihre Kompatibilität mit dem Hardy-Weinberg Gleichgewicht kontrolliert.
Die Tests auf Abweichungen vom Hardy-Weinberg Gleichgewicht sind nur als ein Anhalt für mögliche Genotypisierungsfehler zu sehen. Wie in Abschnitt 9.1 beschrieben, sind Abweichungen nicht nur durch Fehler möglich. Im Gegensatz zu den Mendel-Fehlern kann man bei diesen Tests nicht immer sicher sagen, ob Genotypisierungsfehler vorliegen oder nicht. Die gefundenen Abweichungen müssen schon sehr groß sein, um sicher als Folge von Genotypisierungsfehlern interpretiert werden zu können.
In der Auswertung im Rahmen dieser Arbeit werden alle Marker als „auffällig“ bezeichnet, die beim χ^2-Test und bei der Monte-Carlo Permutation p-Werte von ≤ 0.05 aufweisen. Bei dem Test auf das Vorliegen von Nullallelen werden ebenfalls p-Werte von ≤ 0.05 und ein Wert von ≥ 0.02 für den ML-Schätzer als Kriterium für auffällige Marker benutzt.

Werden „auffällige“ Marker gefunden, hat man die Möglichkeit, die Genotypisierung für diese Marker zu wiederholen oder die Marker aus der Analyse auszuschließen.
Im Rahmen der LRS-Studie werden Marker, die in den Tests zum Hardy-Weinberg Gleichgewicht auffällig sind, besonders beobachtet. Ergeben sich auffällige p-Werte in den Auswertungen, wird die Analyse einmal mit und einmal ohne die betreffenden Marker

durchgeführt. Ergeben sich genau an der Stelle, an der die Marker stehen, auffällige Ergebnisse, werden die Marker erneut kontrolliert und eventuell ausgeschlossen.

Im Kapitel 11.4 ist beschrieben, welche auffälligen Ergebnisse die Auswertungen unter Nutzung des Hardy-Weinberg Gleichgewichtes ergaben, und wie damit im speziellen Fall umgegangen wird.

10 Standardarbeitsanweisung zur Qualitätssicherung

Ziel der vorliegenden Arbeit ist es, Empfehlungen zur Durchführung einer Qualitätssicherung bei Genotypisierungsdaten zu geben. Die Kontrollen der Genotypisierungsdaten, die im Rahmen der LRS-Studie im IMBS durchgeführt wurden, sind übertragbar auf andere Studien ähnlichen Designs, das heißt, familienbasierte Studien mit Mikrosatelliten. In diesem Kapitel werden die Methoden zur Qualitätssicherung in Form einer Standardarbeitsanweisung (engl. *Standard operating procedure*, SOP) dargestellt. Die Idee hinter einer solchen SOP ist, dass Prozesse innerhalb eines Institutes oder eines Unternehmens in immer gleicher Weise ablaufen. Eine SOP ist so genau, dass zwei verschiedene Bearbeiter für ein Problem mit denselben Schritten zu denselben Ergebnissen kommen. Diese Ergebnisse müssen immer nachvollziehbar sein.

10.1 Einleitung

Diese SOP beinhaltet Richtlinien zur Qualitätssicherung von Genotypisierungsdaten in familienbasierten Studien mit Mikrosatelliten. Allgemeine Hintergründe zu Fehlern in molekulargenetischen Studien sind im Kapitel 4 dieser Arbeit beschrieben. In den Kapiteln 5 bis 9 wurden verschiedene Methoden beschrieben, mit denen Genotypisierungsfehler in Daten gefunden werden können.

In dieser SOP wird die praktische Umsetzung der Methoden beschrieben, und es wird erklärt, wie man mit entdeckten Fehlern umgeht. Diese SOP gliedert sich in zehn Abschnitte. Die Abschnitte Software (10.2) und Datenmanagement (10.3) enthalten allgemeine Hinweise, die anschließenden Abschnitte erläutern jeweils die einzelnen Schritte der Qualitätskontrolle.

Zielgruppe dieser SOP sind mit der Durchführung der Qualitätskontrollen betraute Personen. Sie werden im Folgenden als Bearbeiter bezeichnet. Die Bearbeitung der Daten erfolgt unter Supervision des verantwortlichen Biometrikers. Fragen zu Zugriffe auf benötigte Dateien, zu Dateiformaten, Programmen und Ähnlichem werden mit dem verantwortlichen Administrator geklärt.

10.2 Software

Zur Kontrolle auf Genotypisierungsfehler werden verschiedene Programme benutzt. Die Kontrolle auf Mendel-Fehler wird mit dem Programm *Pedcheck*, Version 1.0, (O'Connell & Weeks, 1998) die Suche nach Doppelrekombinanten wird mit dem Programm *Genehunter*, Version 2.1r_5beta, (Kruglyak et al., 1996) durchgeführt. Die Kartierung erfolgt mit dem Programm *Crimap*, Version 2.4 (Green et al., 1990). Ein Teil der Tests, die das Hardy-Weinberg Gleichgewicht prüfen, werden mit den Programmen *Pedstats 0.5.4* (Wigginton & Abecasis, 2005) und *NullAlleleCheck*, Version 0.98, (2004) durchgeführt. Alle fünf der oben genannten Programme werden auf *Unix* verwendet.
Die Permutation zum Hardy-Weinberg Gleichgewicht ist als Algorithmus in *R*, Version 2.2.1, (2005) implementiert. Die Software *R* wird auf *Windows* verwendet.
Die Bearbeitung der Dateien erfolgt unter *Windows* zum Beispiel mit dem Programm *UltraEdit*. Die Dateien müssen so abgelegt werden, dass man von *Windows* und *Unix* auf sie zugreifen kann. Im IMBS wird dies mit dem verantwortlichen Administrator abgeklärt.

10.3 Datenmanagement

Die Genotypisierungsdaten liegen in Form einer *.ped* und einer *.dat* Datei vor. Diese Formate sind im Kapitel 5 dieser Arbeit beschrieben. Hier werden sie als *ihredaten.ped* und *ihredaten.dat* bezeichnet.
Für jedes Chromosom liegen je eine Datei *ihredaten.ped* und *ihredaten.dat* vor. Die Schritte der Qualitätssicherung werden dementsprechend jeweils für jedes Chromosom durchgeführt.
Während der Durchführung der Qualitätssicherung werden Daten vor allem in der Datei *ihredaten.ped* modifiziert, wenn Fehler gefunden werden. Es wird dokumentiert, welche Personen für welche Marker modifiziert werden. Dies kann zum Beispiel in einer *Excel*-Tabelle erfolgen. Ergibt eine Kontrolle keine Fehler und werden keine Modifikationen durchgeführt, wird dies ebenfalls protokolliert.
Im Kopf der Tabelle werden eindeutig die angewandte Methode (also z.B. *Pedcheck*) sowie Datum und Uhrzeit der Bearbeitung angegeben. Führen verschiedene Bearbeiter die Qualitätssicherung durch, wird jeweils der Name desjenigen angegeben, der eine Kontrolle gemacht hat. In der Tabelle sind die Identifikation der Person in der Studie, die Bezeichnung

des betroffenen Markers sowie die originalen und die neuen (also modifizierten) Allele enthalten.

Von der Originaldatei *ihredaten.ped* wird mindestens eine Kopie erstellt, die in der Qualitätssicherung nicht modifiziert wird. Treten während der Qualitätssicherung Fehler auf oder gehen Daten verloren, kann auf diese Kopie zurückgegriffen werden. Die Originaldatei wird getrennt von den bearbeiteten Dateien gespeichert, zum Beispiel als *ihredaten.ped_ORIGINAL*.

Nach Abschluss einer Methode zur Entdeckung von Genotypisierungsfehlern, wird ebenso mit der modifizierten Datei verfahren; eine Kopie wird gespeichert, gesondert abgelegt und im Folgenden unverändert gelassen. Mit einer anderen Kopie wird der nächste Schritt der Qualitätssicherung durchgeführt.

10.4 Suche nach Mendel-Fehlern mit *Pedcheck*

Als erster Schritt der Qualitätssicherung werden die Genotypisierungsdaten mit dem Programm *Pedcheck* geprüft. Der Algorithmus hierzu ist in Kapitel 6 beschrieben.

Das Programm *Pedcheck* läuft auf einer *Unix*-Oberfläche. Die benötigten Dateien *ihredaten.ped* und *ihredaten.dat* müssen auf *Unix* vorliegen.

Unter *Unix* wird die erste Stufe der Prüfung mit *Pedcheck* durchgeführt.

Die von *Pedcheck* entdeckten Fehler werden vom Programm nach jedem Programmdurchlauf in der Datei *pedcheck.err* gespeichert. Diese Zusammenfassung der Fehler kann auf der *Unix*-Ebene betrachtet (Befehl: *more pedcheck.err*) oder *pedcheck.err* kann mit *UltraEdit* unter *Windows* geöffnet werden.

Meldet *Pedcheck* Fehler, so wird die Datei *ihredaten.ped* mit *UltraEdit* geöffnet. Je nachdem was für ein Fehler vorliegt, werden die Genotypen so modifiziert, dass der Fehler korrigiert ist.

In den meisten Fällen werden nur bei einer Person für einen Marker die Allele modifiziert werden müssen. Weist zum Beispiel ein Kind in einer Kernfamilie fehlerhafte Allele für einen bestimmten Marker auf, werden die Originalallele, die *Pedcheck* als fehlerhaft erkannt hat, bei dem Kind „genullt“, das heißt, durch die neuen Allele (0/0) ersetzt.

Es sind aber auch komplizierte Fälle möglich. Tritt zwischen zwei Personen Inkonsistenz auf, muss überlegt werden, welche der zwei Personen fehlerhaft genotypisiert worden ist. Dann

können entweder nur die betroffenen Allele dieser einen Person oder – falls Unklarheit besteht – beider Personen genullt werden.

Es kann der Fall auftreten, dass eine Inkonsistenz durch andere Modifizierung der Allele außer dem „Nullen“ behoben werden kann. Ein Beispiel wäre die Person 3 im Beispielstammbaum aus der Abbildung 6.1.

Hier können die von *Pedcheck* gefundenen Fehler behoben werden, indem die Person 3 die neuen Allele (3/1) erhält. Eine derartige Modifikation setzt aber voraus, dass man sich sicher ist, dass die Daten bei den anderen Personen im Stammbaum alle richtig sind. Diese Sicherheit besteht aber nicht. Ebenso könnten in diesem Beispiel die Person 7 und/oder die Personen 4 oder 5 falsch genotypisiert sein. Würde man in diesem Beispiel die Daten bei der Person 3 anders als nach (0/0) modifizieren, vergrößert man eventuell sogar einen Genotypisierungsfehler.

Durch „Nullen“ der als fehlerhaft erkannten Allele, verliert man unter Umständen zwar Informationen, aber es können keine zusätzlichen Fehler in den Daten erzeugt werden.

Die Datei *ihredaten.ped* wird gespeichert und *UltraEdit* verlassen. Die *Pedcheck* Kontrolle wird wiederholt, und wenn keine weiteren Fehler auf der geprüften Ebene auftreten, wird die nächste *Pedcheck* Ebene ausgeführt.

Tabelle 10.1 zeigt die Kontrolle der Genotypisierungsdaten im chronologischen Ablauf. Die Aktionen sind je nach benutzter Oberfläche mit *Unix* oder *UltraEdit* betitelt.

1.	*Unix* Geben Sie den Befehl `pedcheck -p ihredaten.ped –d ihredaten.dat`, um die erste Stufe von *Pedcheck* durchzuführen. Die Ausgabe der Fehler erfolgt in der Datei `pedcheck.err`. Sie kann unter *Unix* (Befehl: `more pedcheck.err`) oder unter *Windows* mit *UltraEdit* betrachtet werden.
2.	*UltraEdit* Korrektur der fehlerhaften Genotypen in der Datei `ihredaten.ped` (siehe Text). Dokumentation der Modifizierungen. Speichern der Datei. Wiederholung von (1) und (2) solange, bis keine Fehler mehr gemeldet werden.

3.	*Unix* Geben Sie den Befehl: `pedcheck -2 -p ihredaten.ped –d ihredaten.dat`, um die zweite Stufe von *Pedcheck* durchzuführen. Die Ausgabe der Fehler erfolgt in der Datei `pedcheck.err`.
4.	*UltraEdit* Korrektur der fehlerhaften Genotypen in der Datei `ihredaten.ped`. Dokumentation der Modifizierungen. Speichern der Datei. Wiederholung von (3) und (4) solange, bis keine Fehler mehr gemeldet werden.
5.	*Unix* Geben Sie den Befehl `pedcheck -3 -p ihredaten.ped –d ihredaten.dat`, um die dritte Stufe von *Pedcheck* durchzuführen. Die Ausgabe der Fehler erfolgt in der Datei `pedcheck.err`.
6.	*UltraEdit* Korrektur der fehlerhaften Genotypen in der Datei `ihredaten.ped`. Dokumentation der Modifizierungen. Speichern der Datei. Wiederholung von (5) und (6) solange, bis keine Fehler mehr gemeldet werden.
7.	*Unix* Geben Sie den Befehl `pedcheck -4 -p ihredaten.ped –d ihredaten.dat`, um die vierte Stufe von *Pedcheck* durchzuführen. Die Ausgabe der Fehler erfolgt in der Datei `pedcheck.err`.
8.	*UltraEdit* Korrektur der fehlerhaften Genotypen in der Datei `ihredaten.ped`. Dokumentation der Modifizierungen. Speichern der Datei. Wiederholung von (7) und (8) solange, bis keine Fehler mehr gemeldet werden.

Tabelle 10.1
Übersicht zur Prüfung der Daten mit *Pedcheck*

Die Datei *pedcheck.err* wird bei jedem neuen Programmdurchlauf automatisch überschrieben. Sollen die Fehlermeldungen aus *pedcheck.err* nach einem Durchlauf gespeichert werden, muss der Inhalt von *pedcheck.err* in einer Datei mit einem anderen Namen abgelegt werden.

Wie im Abschnitt 10.3 beschrieben, wird festgehalten, welche Personen für welche Marker modifiziert sind. Wenn *Pedcheck* abgeschlossen ist, wird die Datei *ihredaten.ped* gespeichert. Eine Kopie wird gesondert abgelegt (z.B. als *ihredaten.ped_MENDELKORREKT*) und im weiteren Verlauf der Qualitätssicherung nicht mehr modifiziert. Eine andere Kopie wird im nächsten Schritt der Qualitätssicherung weiter bearbeitet.

Während und nach der Durchführung des Mendel-Checks wird dokumentiert, wie viele Mendel-Fehler auftreten. Ein Maß sind etwa die ausgeschlossenen Loci bezogen auf alle Loci. Diese Fehlerquote wird dokumentiert (z.B. auch im Kopf in der *Excel*-Datei). Eine Quote von unter einem Prozent kann akzeptiert werden. Eine höhere Quote muss vom Bearbeiter sofort an den verantwortlichen Biometriker gemeldet werden. Treten innerhalb einzelner Familien oder bei bestimmten Markern gehäuft Fehler auf, wird auch das dokumentiert und gemeldet.
Der Biometriker entscheidet in so einem Fall über das weitere Vorgehen.

10.5 Ausschluss von Doppelrekombinationen mit *Genehunter*

Nach abgeschlossener Prüfung auf Mendel-Fehler werden die Daten mit dem Programm *Genehunter* auf das Vorliegen von Doppelrekombinanten geprüft. Im Kapitel 7 dieser Arbeit ist dieser Schritt der Qualitätssicherung beschrieben.
Genehunter wird wie *Pedcheck* auf *Unix* verwendet. Die Daten müssen als eine *.ped* Datei und eine *.dat* vorliegen. Die zuvor im *Pedcheck* modifizierten Dateien *ihredaten.ped* und *ihredaten.dat* werden direkt weiter bearbeitet.
Das Programm *Genehunter* bestimmt für jede Person in allen Stammbäumen die wahrscheinlichsten Haplotypen. Die Haplotypen werden in einer *postscript* Datei gespeichert. Diese Datei wird ausgedruckt oder am Bildschirm betrachtet. Abbildung 7.3 zeigt eine solche *postscript* Datei. Wenn bei einer Person eine Rekombination auftritt, ist diese durch ein „x“ neben dem Haplotyp gekennzeichnet. Treten zwei Rekombinationen unmittelbar nebeneinander auf, wird vom Vorliegen einer Doppelrekombination ausgegangen. In Abbildung 7.3 ist dies bei der Person 1021-0021696 der Fall.
Wie im Kapitel 7 beschrieben, lässt sich nicht sicher sagen, wo die Ursache der Doppelrekombination liegt. Aus diesem Grund wird bei Auftreten einer Doppelrekombination bei einem Familienmitglied die gesamte Familie für den betroffenen Marker genullt. Wie im vorherigen Schritt erfolgt die Modifizierung der Allele zu (0/0) in *UltraEdit*.

Die Tabelle 10.2 gibt einen Überblick über die Kontrolle auf das Vorliegen von Doppelrekombinanten mit *Genehunter*. Es ist jeweils angegeben, ob ein Arbeitsschritt auf *Unix* oder auf *Windows* mit *UltraEdit* oder *Ghostview* erfolgt.

1.	*Unix* Geben Sie den Befehl `gh`.
2.	Geben Sie den Befehl `ps on`, um die *postscript* Funktion einzuschalten.
3.	Geben Sie den Befehl `haplo on`, um die *haplotype* Funktion einzuschalten.
4.	Geben Sie den Befehl `load marker ihredaten.dat`.
5.	Geben Sie den Befehl `scan pedigree ihredaten.ped`.
6.	Genehunter fragt nun, unter welchem Dateinamen Sie die *postscript out* Datei für die erste Familie im *.ped* Daten speichern möchten: `file to store pedigree plot [F001.ps]:` Sie können einen gewünschten Dateinamen für die *.ps* Datei der ersten Familie eingeben oder durch Drücken von Return die Datei unter *F001.ps* speichern. Der Schritt 6 wiederholt sich für alle Familien in der *.ped* Datei.
7.	*Ghostview* Die Datei *F001.ps* wird mit dem Programm *Ghostview (gsview32)* geöffnet. Alle Haplotypen der Familie werden auf unmittelbar benachbarte Doppelrekombinationen kontrolliert.
8.	*UltraEdit* Liegt eine Doppelrekombination vor, werden die Allele des betroffenen Markers in *ihredaten.ped* bei allen Familienmitgliedern gleich (0/0) gesetzt. Die Modifizierung wird dokumentiert. Die Schritte 7 und 8 werden für alle Familien durchgeführt.

Tabelle 10.2
Übersicht zur Kontrolle auf Doppelrekombinationen mit *Genehunter*

Die modifizierten Dateien *ihredaten.ped* und *ihredaten.dat* werden wie zuvor doppelt gespeichert. Eine Kopie wird zum Beispiel als *ihredaten.ped_DOPPEL* und *ihredaten.dat_DOPPEL* bezeichnet gesondert abgespeichert und im Folgenden unverändert gelassen.

Während und nach der Durchführung der Kontrolle auf Doppelrekombinationen erfolgt eine Dokumentation. Auffälligkeiten wie hohe Quoten an Doppelrekombinanten für bestimmte Familien oder Marker werden vom Bearbeiter an den verantwortlichen Biometriker gemeldet.

10.6 Kartierung der Marker mit *Crimap*

Im Kapitel 8 wurde beschrieben wie mit dem Programm *Crimap* die Kartierung der Marker erfolgen kann. Dort findet sich auch eine Beschreibung der benötigten Input-Dateien. Im IMBS wird mit dem Administrator besprochen, wie die vorhandenen Dateien *ihredaten.ped* und *ihredaten.dat* in *Crimap* kompatible Dateien überführt werden können. *Crimap* wird auf *Unix* ausgeführt. Die Output-Dateien können mit *UltraEdit* betrachtet und gegebenenfalls bearbeitet werden.

Unter Verwendung von *Crimap* wird für jedes Chromosom eine Karte mit den in der Studie verwendeten Markern erstellt. Die erstellte Karte wird mit Referenzkartierungen verglichen, wobei drei Aspekte geprüft werden. Erstens müssen die Chromosomenlängen in der erstellten Karte in etwa mit den Chromosomenlängen in den Referenzkartierungen übereinstimmen. Zweitens wird kontrolliert, ob sich die Reihenfolge der Marker in den Referenzkartierungen nachvollziehen lässt. Und drittens wird geprüft, ob alle für ein Chromosom kartierten Marker auch wirklich auf dem entsprechenden Chromosom liegen.

Mit dem verantwortlichen Biometriker wird abgesprochen, welche Kartierungen als Referenz gewählt werden.

Die Durchführung der Kartierung wird dokumentiert. Die Output-Dateien der Kartierungen werden gespeichert.

Große Abweichungen bei den Chromosomenlängen, falsche Reihenfolge der Marker oder Marker, die in den Referenzkartierungen nicht dem entsprechenden Chromosom zugeordnet werden können, werden an den verantwortlichen Biometriker gemeldet.

Der verantwortliche Biometriker entscheidet bei Auftreten von Fehlern oder Unregelmäßigkeiten über das weitere Vorgehen.

Tabelle 10.3 gibt einen Überblick über die Arbeitsschritte, die während der Kartierung durchgeführt werden.

1.	*Unix* Die benötigten Dateiformate müssen auf *Unix* vorliegen. Geben Sie den Befehl `crimap ihredaten build > build2ihredaten.out`. *Crimap* erstellt eine Kartierung mit den Markern aus ihrem Datensatz.
2.	*UltraEdit* Kontrollieren Sie die erstellte Kartierung unter Verwendung einer Referenzkartierung. Folgende Bedingungen soll die Kartierung erfüllen: - Chromosomenlängen stimmen in ihrer und in der Referenzkartierung in etwa überein. - Die Reihenfolge der Marker lässt sich in der Referenzkartierung nachvollziehen. - Alle kartierten Markern liegen auf dem entsprechenden Chromosom. Die Kontrolle wird dokumentiert. Die Schritte 1 und 2 werden jeweils für jedes Chromosom durchgeführt.

Tabelle 10.3
Schritte zur Qualitätssicherung während der Kartierung mit *Crimap*

10.7 Hardy-Weinberg Gleichgewicht mit *Pedstats*

Die Qualitätssicherung von Genotypisierungsdaten wird mit zwei Tests zum Hardy-Weinberg Gleichgewicht fortgeführt, die im Kapitel 9 dieser Arbeit vorgestellt wurden. Der erste Test ist ein χ^2-Test, der mit dem Programm *Pedstats* durchgeführt wird. Der zweite Test beruht auf einer Monte-Carlo Permutation, die in Abschnitt 10.8 beschrieben wird.
Pedstats wird auf *Unix* verwendet. Die benötigten Input-Dateien haben das *.ped* und das *.dat* Format, die Dateien *ihredaten.ped* und *ihredaten.dat* werden ohne weitere Modifikationen bearbeitet.
Mit dem Programm *Pedstats* werden für jedes Chromosom alle Founder auf Kompatibilität mit dem Hardy-Weinberg Gleichgewicht geprüft. Die Durchführung der Überprüfung sowie auffällige Marker, das heißt, Marker mit p-Werten kleiner 0.05, werden dokumentiert.
Die Ergebnisse des Tests mit *Pedstats* alleine reichen nicht aus, um Marker auszuschließen. Sie werden zusammen mit den Ergebnissen der Monte-Carlo Permutation betrachtet.
Tabelle 10.4 zeigt die Schritte der Qualitätssicherung unter Nutzung von *Pedstats* im chronologischen Ablauf.

1.	*Unix* Geben Sie den Befehl `pedstats -p ihredaten.ped -d ihredaten.dat --hardyweinberg –checkfounders`. *Pedstats* gibt die Marker zurück, die p-Werte kleiner als 0.05 aufweisen.
2.	Dokumentieren Sie auffällige Marker. Wiederholen Sie die Schritte 1 und 2 für alle Chromosomen.

Tabelle 10.4
Schritte zur Kontrolle der Daten mit *Pedstats*

10.8 Monte-Carlo Permutation zum Hardy-Weinberg Gleichgewicht

Die Monte-Carlo Permutation ist als Algorithmus in *R* implementiert (Nutzung in Absprache mit dem Administrator). Das Programm *R* wird unter *Windows* verwendet. Als Input-Datei wird eine Datei vom *.ped* Format benötigt. Es wird eine Kopie der Datei *ihredaten.ped* verwendet, in der die *Non-Founder*, also Kinder, nicht enthalten sind.
Der Algorithmus liegt als Datei im *UltraEdit* Format vor. Zur Durchführung wird das Programm *R* gestartet und der Algorithmus eingefügt (Kopierfunktion *STRG-C* und *STRG-V*). Im Algorithmus wird die Quelle modifiziert, unter der *R* die jeweilige Datei *ihredaten.ped* findet. Die Zahl der Permutationen kann variiert werden. Als Ergebnis wird die Anzahl der beobachteten Homozygoten im Datensatz und einen Monte-Carlo p-Wert zurückgegeben. Optional kann auch ein Plot erstellt werden wie in den Abbildungen 9.3 bis 9.5 dargestellt.
Die Anzahl der beobachteten Homozygoten und der Monte-Carlo p-Wert werden dokumentiert.

Die Ergebnisse der Auswertung mit *Pedstats* und der Auswertung mit der Monte-Carlo Permutation werden in eine Tabelle übertragen (siehe Abbildung 11.4). Fallen in der Zusammenschau beider Methoden Marker mit niedrigen p-Werten auf, wird dies vom Bearbeiter an den verantwortlichen Biometriker gemeldet. Treten Marker mit auffälligen p-Werten auf, wird die Kopplungsanalyse zweimal durchgeführt. Einmal wird die Auswertung mit den auffälligen Markern und einmal ohne sie durchgeführt. Ergeben die Analysen interessante Ergebnisse, das heißt, Hinweise auf Kopplung oder ähnliches, entscheidet der

verantwortliche Biometriker, wie weiter vorgegangen werden soll. Ergeben sich für die auffälligen Marker keine interessanten Ergebnisse, können die Marker in der Analyse belassen werden.
Die Durchführung der Monte-Carlo Permutation sowie ihre Ergebnisse werden dokumentiert.

10.9 Suche nach Nullallelen

Die Suche nach Nullallelen wird mit dem Programm *NullAlleleCheck* durchgeführt, wie im Abschnitt 9.6 dieser Arbeit beschrieben. Das Programm läuft unter Unix und benötigt als Input-Dateien die *.ped* und *.dat* Datei. *ihredaten.ped* und *ihredaten.dat* können also direkt geprüft werden.
Die Ausgabe erfolgt auf der Unix-Oberfläche in Form von Spalten. Die erste Spalte enthält den Marker-Namen, dann die Gesamtzahl der untersuchten Personen, die beobachtete und erwartete Anzahl an Homozygoten, den berechneten χ^2-Wert und p-Wert sowie den Wert $\hat{p}_x$, also den Wert des ML-Schätzers für die Nullallelfrequenz.
Die Ausgabe von *NullAlleleCheck* wird kopiert, in eine *UltraEdit* Datei übertragen und gespeichert. In der Zusammenschau mit den Ergebnissen der Auswertung mit *Pedstats* und der Monte-Carlo Permutation werden die p-Werte und die $\hat{p}_x$-Werte kontrolliert. Auffällige Werte, dass heißt, p-Werte unter 0.05 und $\hat{p}_x$-Werte über 0.02 werden an den verantwortlichen Biometriker gemeldet.
Der Biometriker entscheidet in diesem Fall über das weitere Vorgehen.
Die Durchführung von *Pedstats* wird dokumentiert.

Tabelle 10.5 zeigt die Auswertung mit *NullAlleleCheck* im chronologischen Ablauf.

1.	*Unix* Geben Sie den Befehl `nullac -p ihredaten.ped -d ihredaten.dat`. *NullAlleleCheck* berechnet die Daten und gibt als Ergebnis sieben Spalten zurück: `Marker N Hom(Obs) Hom(Exp) X^2 p-Value pHatX` … … … … … … …
2.	Kopieren Sie die Spalten.

3.	*UltraEdit* Speichern Sie die Ergebnisse unter UltraEdit. Wiederholen Sie die Schritte 1 bis 3 für alle Chromosomen.
4.	Kontrollieren Sie Ergebnisse auf auffällige p-Werte und $\hat{p}_x$ -Werte. Dokumentieren Sie die gesamte Auswertung.

Tabelle 10.5
Durchführung der Qualitätssicherung unter Verwendung von *NullAlleleCheck* im Überblick

10.10 Kontrollen auf fehlerhafte Relationen und Geschlechtsangaben

Im Kapitel 4 wurden Stammbaumfehler beschrieben. Als Stammbaumfehler bezeichnet man falsche Relationen der Personen in den Genotypisierungsdaten. Nach diesen Fehler wird mit den Programmen *PREST* und *ALTERTEST* gesucht (McPeek & Sun, 2000; Sun et al., 2002), die in dieser Arbeit nicht behandelt wurden.
Ein den Stammbaumfehlern ähnlicher Fehler liegt vor, wenn für eine Person ein falsches Geschlecht in den Genotypisierungsdaten angegeben ist. Dieser Fehler wird ausgeschlossen, indem das in den Genotypisierungsdaten erfasste Geschlecht mit dem Geschlecht in den Daten verglichen wird, die am Anfang der Studie bei der Erfassung der Personen erhoben wurden.
Die Durchführung der Kontrollen auf Stammbaumfehler oder Fehler bei Geschlechtsangaben in den Genotypisierungen wird dokumentiert. Liegen Fehler vor, werden sie dokumentiert und an den verantwortlichen Biometriker gemeldet.

11 Ergebnisse der Anwendungen auf die Daten aus der Dyslexie-Studie

In diesem Kapitel sind die Ergebnisse zusammengestellt, die die verschiedenen Methoden zur Qualitätskontrolle angewandt auf die Genotypisierungsdaten der LRS-Studie ergaben. Im Abschnitt 11.5 sind die Ergebnisse zusammengefasst.
Eine Diskussion der Ergebnisse und Vorschläge zur Kontrolle auf Genotypisierungsfehler bei familienbasierten Studien mit STR-Markern findet im Kapitel 12 statt.

11.1 Mendel-Fehler

Der Test auf Mendel-Fehler fand wie in den Kapiteln 6 und 10 beschrieben mit Hilfe des Programms *Pedcheck* statt.
Die Kontrolle auf Mendel-Fehler führte zum Ausschluss von 557 Loci bei den jeweiligen Personen, das sind 0.11% aller genotypisierten Loci. Die ausgeschlossenen Loci wurden bei den betroffenen Personen auf „0 0“ gesetzt. Die Tabelle 11.1 enthält einen Auszug aus den ausgeschlossenen Loci (identifiziert durch Markername und Personen ID) mit ihren Original Allelen.

Marker	Person ID	Original Allele	
D1S205	1021-0021522	117	117
D1S205	1021-0021523	106	106
D1S205	1021-0021524	113	117
D1S205	1021-0021525	113	113
D1S205	1021-0022072	111	113
D1S237	1021-0022007	188	198
D1S249	1021-0023943	165	175
D1S2620	1021-0021819	103	107
D1S2628	1021-0021792	121	125
D1S2652	1021-0021362	105	105
D1S2652	1021-0021363	103	103
D1S2667	1021-0023975	136	160
D1S2788	1021-0022059	220	222
D1S2788	1021-0023917	212	222
D1S2797	1021-0023970	121	123
D1S2870	1021-0021792	208	212
D1S430	1021-0022165	187	191
D1S434	1021-0021451	142	142
D1S434	1021-0021452	142	142
D1S434	1021-0021454	142	142
D1S434	1021-0021455	142	142
D1S434	1021-0021639	141	142

Marker	Person ID	Original Allele	
D1S434	1021-0022035	141	142
D1S450	1021-0021464	337	343
D1S450	1021-0021465	337	347
D1S450	1021-0021466	337	349
D1S452	1021-0021825	128	130
D1S452	1021-0021826	124	132
D1S452	1021-0021827	130	132
D1S452	1021-0021828	124	130
D1S452	1021-0021829	130	132
D1S484	1021-0022116	276	280
D1S484	1021-0022118	282	282
D1S484	1021-0022119	278	282
D1S495	1021-0021656	155	155

Tabelle 11.1

Auszug aus Ergebnissen der Mendel-Checks für Chromosom 1

11.2 Doppelrekombinationen

Die Kontrolle auf Doppelrekombinationen erfolgte wie in den Kapiteln 7 und 10 erklärt. Unter Nutzung des Programms *Genehunter* wurden die wahrscheinlichsten Haplotypen für alle Familien für jedes

Chrom.	Familie	Marker-Position	Marker
Chr1	F007	34	D1S2818
Chr1	F010	34	D1S2818
Chr1	F011	15	D1S476
Chr1	F027	27	D1S189
Chr1	F032	34	D1S2818
Chr1	F037	6	D1S2697
Chr1	F037	7	D1S2620
Chr1	F041	7	D1S2620
Chr1	F044	7	D1S2620
Chr1	F084	35	D1S413
Chr1	F094	6	D1S2697
Chr1	F099	9	D1S255
Chr1	F111	41	D1S237
Chr1	F113	45	D1S2850

Tabelle 11.2

Von Doppelrekombinationen betroffenen Loci auf Chromosom 1

Autosom rekonstruiert und auf Doppelrekombinanten inspiziert. Bei Personen, bei denen sich Doppelrekombinanten zeigten, wurden die Allele für die betroffenen Marker in der ganzen Familie auf „0 0“ gesetzt.
Insgesamt wurden so in betroffenen Familien 1017 Loci (0.19%) aus der Analyse genommen.
Die Tabelle 11.2 gibt für das Chromosom 1 die betroffenen Marker und Familien wieder.

11.3 Kartierung der Marker

Die Kartierung der Marker ist in den Kapiteln 8 und 10 beschrieben. Bei der Kartierung ergaben sich keine Auffälligkeiten. Es wurden sieben Marker aus dem Datensatz genommen, weil bei ihnen weniger als 50% der Rohgenotypisierungen erfolgreich waren. Hierdurch fielen 2363 Loci (0.45%) aus der Auswertung.
Die Tabelle 11.3 enthält die ausgeschlossenen Marker mit der jeweiligen Anzahl ausgeschlossener Loci.

Marker	Anzahl ausgeschlossene Loci
D1S305	78
D1S197	93
D3S1607	266
D6S1637	268
D7S2201	300
D8S1707	529
D9S1793	829
Gesamt	2363

Tabelle 11.3
Ausgeschlossene Marker

11.4 Tests unter Verwendung des Hardy-Weinberg Gleichgewicht

Die Überprüfung auf Abweichung von der Hardy-Weinberg Verteilung wurde für alle Marker bei den *Foundern* – also den Eltern in den Stammbäumen – durchgeführt. Die Durchführung dieser Überprüfung ist in den Kapiteln 9 und 10 erläutert.
Es wurden keine Loci aus der Analyse ausgeschlossen.

Die Kontrolle auf Übereinstimmungen der Genotypisierungsdaten mit der Hardy-Weinberg Verteilung wurde in mehreren Schritten ausgeführt. Es wurde ein χ^2-Test unter Verwendung des Programm *Pedstats* durchgeführt, eine Monte-Carlo Permutation und eine Kontrolle auf Nullallele mit dem Programm *NullAlleleCheck*. Die graphische Kontrolle der Marker mit Hilfe der Vasarely Methode wurde für einige Chromosomen durchgeführt, um die Methode zu testen. Ergebnisse, die den Ausschluss eines Markers aus der Analyse zur Folge hatten, ergab die Auswertung der Vasarely-Tafeln nicht.

Unter Verwendung des χ^2-Tests ergaben sich für die 17 in Tabelle 11.4 aufgeführten Marker auffallende Werte. Mit aufgeführt sind die Ergebnisse der Monte-Carlo Permutation für diese Marker.

Es ist zu erkennen, dass die p-Werte des χ^2-Tests zum Teil deutlich kleiner als $\alpha = 0.05$ sind.

In einem zweiten Schritt wurden die Genotypisierungsdaten mit der Monte-Carlo Permutation untersucht. Hier zeigten vier von 17 Marker, die schon im χ^2-Test aufgefallen waren, ebenfalls signifikante p-Werte. Bei den übrigen 13 Marker fanden sich jedoch keine signifikanten p-Werte in der Monte-Carlo Simulation. Ein Beispiel dafür ist der Marker D3S1270. Der p-Wert der Auswertung mit *Pedstats* ist $p = 0.0078$ und weist somit auf eine erhöhte Anzahl an Homozygoten bei diesem Marker hin. In der Monte-Carlo Permutation ist der p-Wert mit $p = 0.18875$ unauffällig. Betrachtet man die Anzahl der beobachteten Homozygoten fällt auf, worin ein Grund für die abweichenden Ergebnisse beider Tests liegen kann. *Pedstats* beobachtet für diesen Marker 150 Homozygote und 302 Heterozygote, berechnet aber eine Anzahl an zu erwartenden Heterozygoten von 311. In der Monte-Carlo Permutation werden aber nur 140 Homozygote und 312 Heterozygote beobachtet. Diese Zahlen liegen eher in dem Bereich, der für beide Werte nach den Genotypfrequenzen zu erwarten ist. Der p-Wert ist deshalb bei der Monte-Carlo Permutation höher.

Pedstats beobachtet mehr Homozygote und berechnet einen signifikanten p-Wert, weil das Programm, wie in Abschnitt 9.4 erwähnt, Allele mit geringen Frequenzen als ein Allel betrachtet. Ein Nebeneffekt dieses Allelpoolings ist aber offenbar ein künstlicher Homozygotenexzess. Werden wie beim Marker D3S1270 acht Allele zu einem Allel zusammengefasst, erscheinen Genotypen, die vorher mit zwei verschiedenen dieser acht Allele heterozygot waren, nun als homozygot. *Pedstats* beobachtet für diesen Marker 150 Homozygote, es liegen aber tatsächlich nur 140 Homozygote vor, wie in der Auswertung mit der Monte-Carlo Permutation bestimmt.

Marker	Ergebnisse des χ^2-Test							Ergebnisse der Monte-Carlo Permutation	
	B(hom)	B(het)	E(het)	analysierte Allele	gepoolte Allele	Alle Allele des Markers	p-Wert	MC-B(hom)	MC p-Wert N=20000
D1S2713	120	372	342	4	11	1-14	0.0110	104	0.99145
D1S2688	192	299	303	3	2	1-4	0.0120	188	0.30385
D3S1270	150	302	311	4	8	1-11	0.0078	140	0.18875
D3S3551	78	262	246	5	11	1-15	0.0304	34	0.94485
D3S1580	109	367	367	5	9	1-13	0.0459	75	0.3189
D4S1572	90	384	383	6	6	1-11	0.0379	87	0.2293
D4S1615	140	340	356	4	4	1-7	0.0044	134	0.0396
D5S424	130	329	333	4	10	1-13	0.0284	122	0.25285
D5S2090	105	386	394	6	6	1-11	0.0001	101	0.20555
D7S2463	145	341	312	3	10	1-12	0.0327	117	0.9958
D7S2418	140	347	366	5	3	1-7	0.0305	139	0.0079
D8S279	52	446	426	8	7	1-14	0.0467	45	0.9909
D9S1779	186	288	266	4	7	1-10	0.0338	182	0.99775
D10S1218	8	48	59	2	4	1-5	0.0005	434	5,00E-05
D10S537	77	395	394	7	6	1-12	0.0168	72	0.47455
D14S262	190	297	294	4	3	1-6	0.0217	190	0.53355
D18S554	106	359	381	6	6	1-11	0.0032	94	0.01525

Tabelle 11.4

Im χ^2-Test auffällige Marker und ihre Ergebnisse in der Monte-Carlo Permutation

So kann *Pedstats* für manche Marker einen Homozygotenexzess beobachten, den es selbst produziert hat. Ohne die Testung mit der Monte-Carlo Methode und damit der erneuten Berechnung der beobachteten Homozygoten, fällt diese Verzerrung zunächst nicht auf.
Bei den Markern D4S1615, D7S2418 und D18S554 ergaben sich bei beiden Methoden signifikante p-Werte, und somit scheint bei diesen Markern wirklich ein Heterozygotenmangel vorzuliegen.
Ein Sonderfall ist der Marker D10S1218. Hier ergaben sich bei beiden Auswertungen hoch signifikante p-Werte. Dieser Marker hat fünf Allele, von denen eines mit einer Frequenz von p = 0.94014 vorkommt. Die anderen vier Allele weisen Frequenzen von 0.04367, 0.00883, 0.00589 und 0.00147 auf. Die signifikanten p-Werte könnten allein auf dieser ungleichen Verteilung der Allelfrequenzen beruhen.
Wie erwähnt wurden keine Marker, die im χ^2-Test oder in der Monte-Carlo Permutation auffällige p-Werte ergaben, aus der Analyse ausgeschlossen. Jedoch wurde die Kopplungsanalyse einmal mit und einmal ohne die 17 auffälligen Marker durchgeführt. Es ergaben sich keine relevanten Abweichungen in den Ergebnissen. Die betroffenen Marker gehörten nicht zu den Regionen, die in der Dyslexie Studie für Kopplung oder Assoziation auffällig waren. Wenn sich gerade an diesen Stellen Hinweise für Kopplung oder Assoziation ergeben würde, würden sie aber erneut kontrolliert werden und gegebenenfalls aus der Analyse ausgeschlossen werden.

Der Test auf Nullallele wurde für alle Marker durchgeführt. Die Methode zur Überprüfung ist in Kapitel 9.7 beschrieben. Bei zehn Markern ergaben sich auffällige p-Werte und der ML-Schätzer für die Nullallelfrequenz ergab Werte über 0.02.

Marker	N	Hom(Obs)	Hom(Exp)	χ^2	p-Value	p^
D2S305	1002	251	215.898	7.27457	0.00699375	0.0228366
D5S1981	1022	312	277.209	5.99177	0.0143728	0.0239151
D6S273	1020	283	242.092	9.06389	0.00260707	0.0270038
D7S2418	1023	293	248.98	10.2865	0.00134006	0.0292685
D7S1805	935	233	202.269	5.9579	0.0146515	0.0214194
D10S1676	1029	345	314.098	4.3761	0.0364463	0.0220904
D12S374	901	278	248.282	4.90997	0.0267021	0.0232948
D20S1083	1025	434	406.086	3.17777	0.0746464	0.0230711
D21S1904	939	453	426.635	2.98604	0.0839854	0.0264085
D22S1174	1041	329	296.498	4.98182	0.025615	0.0223152

Tabelle 11.5
Auffällige Marker im Test auf Nullallele

Tabelle 11.5 gibt einen Überblick über die auffälligen Marker.
Einer dieser Marker, D7S2418 fiel auch bei den Tests auf Abweichungen von der Hardy-Weinberg Verteilung auf.
Wie bei den Kontrollen zum Hardy-Weinberg Gleichgewicht wurde eine Kopplungsanalyse jeweils mit und ohne die auffälligen zehn Marker durchgeführt. Da sich keine Verschiebung in den Ergebnissen zeigt, und die Marker nicht an relevanten Stellen lagen, wurden sie nicht aus der Analyse ausgeschlossen.
Hätten sich bei den betroffenen Markern Auffälligkeiten in der Kopplungsanalyse ergeben, so würden sie erneut kontrolliert und eventuell ausgeschlossen werden.
Zum Umgang mit den in den Tests auf das Hardy-Weinberg Gleichgewicht auffälligen Markern siehe auch die Diskussion in Kapitel 12.

11.5 Zusammenfassung der Ergebnisse

Die Genotypisierung der Personen aus der Dyslexie Studie fand durch deCode statt. Im Zuge der Genotypisierung und in einem routinemäßig durchgeführten Mendel-Check fielen 6 Personen von ursprünglich 1060 aus dem Datensatz, der an das IMBS gegeben wurde.
Nach der Genotypisierung lagen die Genotypisierungsdaten von 1058 Personen vor. Ursprünglich waren 1060 Personen im Datensatz, aber zwei Personen wurden durch deCode ausgeschlossen, weil sie identische Proben wie je eine andere Person aufwiesen.
Eine ausgeschlossene Person (1021-0021574 aus der Familie F073) war ein nicht-betroffenes Kind – also ein nicht von der Krankheit betroffenes Geschwisterkind. Seine Daten waren identisch zu den Daten seines Vaters (1021-0021573). Der Vater wurde als *Founder* beibehalten, das Kind ausgeschlossen. Das betroffene Geschwisterkind und die Eltern blieben in der Auswertung.
Die zweite ausgeschlossene Person war ein betroffenes Kind (1021-0021834 aus der Familie F133). Es wurde ausgeschlossen, weil seine Daten identisch zu den Daten der Person mit der ID 1021-0021833 (aus der Familie F132) waren. Die Eltern und das nicht-betroffene Geschwisterkind blieben in der Analyse, wurden aber nicht für alle Auswertungen benutzt.

Von den 1058 verbliebenen Personen wurden noch einmal vier Personen ausgeschlossen.
Der Person 1021-0021752 (aus der Familie F114) konnte durch deCode keine Familie zugeordnet werden. Sie wurde ausgeschlossen. Da es sich bei der Person um ein Elternteil der

Familie F114 handelte, musste die ganze Familie F114 (insgesamt vier Personen) aus der Analyse genommen werden.
Die Tabelle 11.6 gibt einen Überblick über die ausgeschlossenen Personen.

Personen ID	Familien ID	Einzel-Person	Familie	Grund für Ausschluss
1021-0021574	F073	X		Gleiche Daten wie andere Person
1021-0021752	F114		X	Keine Zuordnung zu Familie möglich
1021-0021750	F114		X	Folge des Ausschluss von Person 1021-0021752
1021-0021751	F114		X	Folge des Ausschluss von Person 1021-0021752
1021-0021753	F114		X	Folge des Ausschluss von Person 1021-0021752
1021-0021834	F133	X		Gleiche Daten wie andere Person

Tabelle 11.6
Ausgeschlossene Personen

In den Daten, die von deCode an das IMBS gegeben wurden, wo die beschriebenen Methoden zur Qualitätskontrolle stattfanden, enthielten die Genotypen von 1054 Personen für 533 Marker.

Von den 525140 Loci, die erfolgreich genotypisiert worden waren, wurden 3937 Loci in der Qualitätssicherung ausgeschlossen, das sind 0.75%. Aufgrund von Mendel-Fehlern wurden 557 Loci (0.11%) aus der Analyse genommen, 1017 Loci (0.19%) wurden in der Kontrolle auf Doppelrekombinanten ausgeschlossen, und 2363 Loci (0.45%) fielen in der Kartierung aus der Analyse.
Bei der Kontrolle auf Einhaltung des Hardy-Weinberg-Gleichgewichtes ergaben sich bei einigen Markern Auffälligkeiten, sie wurden aber nicht aus der Analyse genommen. In der Prüfung auf Nullallele wurden bei zehn Markern leicht überhöhten Raten an Nullallelen festgestellt. Die Marker wurden ebenfalls nicht ausgeschlossen.
Für alle Chromosomen auf denen auffälligen Marker lagen wurde die Kopplungsanalyse zweimal durchgeführt. Einmal wurden die Marker in der Analyse belassen, einmal wurden sie ausgeschlossen. Da es zu keinen relevanten Verschiebungen in den Ergebnissen kam, und keiner der betroffenen Marker in für Kopplung oder Assoziation interessanten Regionen lag, wurden alle Marker in der Analyse belassen.

Die in dieser Arbeit vorgestellten Methoden zur Entdeckung von Fehlern in Genotypisierungsdaten werden im anschließenden Kapitel diskutiert.

12 Diskussion

Genotypisierungsfehler treten in allen molekulargenetischen Untersuchungen auf, die mit Genotypisierungsdaten arbeiten (Hoffman & Amos, 2005; Pompanon et al., 2005). Ihre vielfältigen Ursachen im menschlichen wie technischen Bereich in allen Schritten einer Untersuchung lassen sich zwar oft nachvollziehen, aber nicht immer verhindern.
Zwei Strategien sollten verfolgt werden, um das Problem der Genotypisierungsfehler zu handhaben. Zum einen sollte versucht werden, die mögliche Entstehung von Genotypisierungsfehlern zu minimieren. Zum anderen sollten die Daten nach der Genotypisierung auf Fehler kontrolliert werden.

Das Studiendesign, eine genaue Erhebung der Daten und präzises Arbeiten im Labor können die Fehlerrate senken. Menschliche Ursachen haben wahrscheinlich den wichtigsten Anteil an der Entstehung von Fehlern (Bonin et al., 2004; Hoffman & Amos, 2005). Darum ist es sinnvoll, die Fehlerquelle Mensch möglichst zu minimieren. Dies kann durch erfahrenes Laborpersonal und möglichst viele standardisierte und validierte Arbeitsschritte erfolgen (Butler, 2001). Wo dies möglich ist, sollten Arbeitsschritte automatisiert werden (Perlin et al., 1995). Genotypisierungsfehler können zum Teil durch den Einsatz von Positiv- und Negativ-Kontrollen in allen Arbeitsschritten weiter minimiert werden. Je nach Fragestellung der Analyse wird empfohlen, 5% bis 10% des Probenmaterials doppelt zu untersuchen. Solche Replikate sollten soweit möglich blind und unabhängig untersucht werden. In manchen Fällen können Genotypisierungsfehler auch dadurch entdeckt werden, dass die Genotypisierung mit einem anderen Verfahren wiederholt wird (Pompanon et al., 2005).
Fehler, die trotz dieser Maßnahmen noch auftreten, können mit verschiedenen Methoden im Nachhinein teilweise entdeckt werden. Wie dies für eine familienbasierte Studie mit STR-Markern aussehen kann, wurde in dieser Arbeit untersucht.

Eine bewährte Methode ist die Nutzung der Familienstruktur, um Fehler zu entdecken, die nicht mit der Mendel'schen Vererbung vereinbar sind. Das Programm *Pedcheck* arbeitet nach diesem Prinzip.
Die entdeckten Fehler bei den Tests auf Mendel-Kompatibilität können durch Fehler in der Genotypisierung entstanden sein. Fast alle dieser Fehler ließen sich dadurch beheben, dass bei den betroffenen Personen die jeweiligen Loci auf Null gesetzt wurden – also als nicht bekannt betrachtet wurden. Dies geschah bei den Mendelfehlern meist nur bei einem Marker bei einer

Person. Das deutet darauf, dass meist nur ein Allel oder ein Allelpaar falsch waren. Diese vereinzelten falschen Allele können aufgrund von technischen Problemen in der Genotypisierung aufgetreten sein. Es kann sich z.B. um die Wiedergabe eines falschen Allels aufgrund von Stotterbanden oder ähnlichem gehandelt haben. Es lag aber nie der Fall vor, dass ganze Haplotypen ausgeschlossen wurden. Dies hätte eher auf mögliche Stammbaumfehler oder vertauschte Proben hingedeutet.
Auch wenn nicht in jedem Fall klar ist, woher der Fehler kommt, können Fehler, die mit den Mendelregeln nicht vereinbar sind, sicher erkannt und ausgeschlossen werden.

Die Interpretation von Doppelrekombinanten als Genotypisierungsfehler ist schon etwas diffiziler. In dieser Arbeit wurde gezeigt, wie mit Hilfe der *haplotype* Funktion des Programm *Genehunter* Familien nach Doppelrekombinanten untersucht werden können. Nachteil der vorgestellten Methode ist, dass die Untersuchung „von Hand“ erfolgt, und somit wiederum der Mensch als potentielle Quelle neuer Fehler auftritt.

Während der Kartierung sollten drei Aspekte beachtet werden, um mögliche Fehler auszuschließen. So sollten die erhaltenen Chromosomenlängen ungefähr mit bekannten Chromosomenlängen übereinstimmen. Die Reihenfolge der Marker sollte so sein wie in Referenzkartierungen. Und für jeden Marker sollte man sicher sein, dass er auch wirklich auf dem entsprechenden Chromosom liegt. Ergeben sich in der Kartierung Auffälligkeiten, sollten sie zusammen mit den Ergebnissen aus den anderen Methoden zur Qualitätssicherung betrachtet werden, um in jedem Fall individuell zu entscheiden, ob Marker aus der Analyse ausgeschlossen werden oder nicht.

Ein wesentlicher Teil dieser Arbeit beschäftigt sich mit der Nutzung des Hardy-Weinberg Gleichgewichtes, um aus Abweichungen von diesem auf mögliche Fehler in den Genotypisierungsdaten zu schließen. Wenn mit dem Gesetz von Hardy und Weinberg gearbeitet wird, sollte stets bedacht werden, dass die Hardy-Weinberg Verteilung nur unter bestimmten Bedingungen gilt. Diese wurden bereits erwähnt: zufällige Paarungen, große Populationen, keine geschlechtsspezifischen Unterschiede in den Allelfrequenzen, Allele segregieren nach den Mendel’schen Gesetzen, kein Wirken evolutionärer Kräfte wie Selektion, Drift, Mutation oder ähnliches.

In der Auswertung der Daten aus der Dyslexie Studie wurde dieser Einschränkung der Hardy-Weinberg Verteilung insofern Rechnung getragen, dass auffällige Ergebnisse der verschiedenen Methoden nicht gleich zum Ausschluss von Markern führten.

Verschiedene Methoden wurden vorgestellt, mit denen Daten auf Abweichungen vom Hardy-Weinberg Gleichgewicht untersucht werden können. Als binominale Methode wurde der χ^2-Test zur Untersuchung eines Heterozygotenmangels verwendet. Unter Verwendung des Programm *Pedstats* können große Datenmengen schnell und einfach kontrolliert werden. Bei der Nutzung dieses Programms muss der *Pooling* Algorithmus berücksichtigt werden und die Ergebnisse dementsprechend kontrolliert werden. Wie gezeigt, verursacht das Programm selbst durch das Zusammenfassen von seltenen Allelen zu einem „neuen" Allel einen gewissen Heterozygotenmangel. Ergeben sich auffällige p-Werte für einen Marker sollte immer geprüft werden, ob eventuell mehrere Allele gepoolt wurden und so eine künstlicher Homozygotenexzess entstanden sein könnte.

Die Monte-Carlo Permutation erwies sich als im Vergleich zum binominalen Heterozygotenmangel-Test als sehr rechenaufwändige Methode. Die im vorher durchgeführten χ^2-Test gefundenen Marker mit auffälligen Abweichungen im Sinne eines Heterozygotenmangels konnten mit der Permutation zuverlässig bestätigt werden. Und auch erst mit der Verwendung der Monte-Carlo Permutation fiel das Problem des von Pedstats selbst erzeugten Homozygotenexzesses auf.

Die Kontrolle auf Nullallele stellt eine weitere Methode dar, Genotypisierungsfehler auszuschließen. Sie erfolgte ebenfalls mit einem χ^2-Test und einer *maximum likelihood* Schätzung nach Fisher (Fisher et al., 2001).

Bei der Kontrolle auf Nullallele fielen zehn Marker auf. Zwei dieser Marker gehörten zu den fünf Markern, die zuvor im χ^2-Test zum Heterozygotenmangel und in der Monte-Carlo Permutation auffielen. Keiner der auffälligen Marker wurde aus den Datensätzen genommen, weil keiner an einem in der späteren Kopplungsanalyse auffallenden Chromosomenabschnitt lag. Zur genaueren Kontrolle wurde die Analyse für Chromosomen mit auffälligen Markern einmal mit und einmal ohne diese Marker durchgeführt.

Sollten sich in anderen Studien gerade für in den Tests auffälligen Markern interessante Ergebnisse in Kopplungsanalysen ergeben, so sollte immer überlegt werden, ob dann die betreffenden Marker verwendet werden können. Die Alternativen wären eine Wiederholung der aufwendigen und teuren Genotypisierung, ein Ausschluss der Marker und damit unter Umständen der Verlust des Ergebnisses oder das Ignorieren der auffälligen Testergebnisse und das in Kauf nehmen von möglicherweise falschen Ergebnissen.

Die Untersuchung der Dyslexie Daten mit Hilfe der Vasarely-Tafeln wurde nicht für alle Chromosomen durchgeführt. Es zeigte sich, dass diese Methode für große Datensätze unpraktikabel ist. Die erhaltenen Ergebnisse sind oft nicht eindeutig zu interpretieren gewesen und unterlagen der subjektiven Wahrnehmung.
Die Vasarely-Methode eignet sich nur für kleine Datensätze mit guter Datenqualität und bekannten Genotyp-Frequenzen. Für eine automatisierte Qualitätskontrolle von großen Datensätzen ist diese Methode nicht zu empfehlen (Ziegler & König, 2006).

Die Zusammenschau der Ergebnisse zur Qualitätskontrolle der Genotypisierungsdaten ergibt bezogen auf die gesamte Anzahl der Loci eine Häufigkeit von 0.75% ausgeschlossenen Loci.
Dieses ist die Zahl der Loci, die mit den hier beschriebenen Methoden nach der Genotypisierung untersucht wurden und auffällig waren. Weil durch deCode schon vorher Loci ausgeschlossen wurden, ist die Zahl der insgesamt ausgeschlossenen Loci etwas höher.

Durch deCode wurden sechs Personen aus der Analyse ausgeschlossen. Von diesen hatten zwei Personen identische Proben mit einer jeweils anderen Person im Studienkollektiv.
Dieser Fehler kann z.B. im Labor durch doppelte Auswertung von Proben oder in einem Schritt der Datenverarbeitung aufgetreten sein. Jedenfalls ist er wahrscheinlich menschlicher Natur.
Die dritte ausgeschlossene Person konnte nicht ihrer Familie zugeordnet werden. Bei dieser Person handelte es sich um ein Elternteil. Deswegen musste die ganze Familie aus dem Datensatz genommen werden. Der Fehler beruhte wahrscheinlich auf einem Etikettierungsproblem – hat also auch eine menschliche Ursache.
Durch die in dieser Arbeit durchgeführten Tests wurden einzelne Loci bei einigen Personen und einige Marker aus der Analyse ausgeschlossen.
Ein Teil der ausgeschlossenen Loci musste aufgrund der Probleme mit den Markern in der Genotypisierung (die Marker, bei denen weniger als 50% der Genotypisierung erfolgreich waren – siehe Kapitel 10.3) aus der Analyse genommen werden. Hierbei handelt es sich nicht direkt um Genotypisierungsfehler, der Verlust dieser Loci ist ja kein Fehler, sondern die Folge eines anderen Fehlers oder einer fehlgeschlagen Genotypisierung aufgrund von Problemen mit den Markern.
Die Kontrolle auf Abweichungen vom Hardy-Weinberg Gleichgewicht und die Kontrolle auf Nullallele führten nicht zum Ausschluss von Loci. Dies kann mit als Beweis der hohen Qualität der Genotypisierung durch das Labor gedeutet werden. Diese Kontrollen wurden

durchgeführt, aber sie sollten und haben keine Auffälligkeiten ergeben, die zum Ausschluss von Genotypisierungsdaten führten.

Diese Arbeit im Rahmen der LRS-Studie befasst sich mit dem Problem, wie für Daten in solchen Genotypisierungsstudien eine „Qualitätssicherung“ aussehen kann.
Der Schwerpunkt lag dabei klar auf den Genotypisierungsfehlern.
Auf die Fehler in Stammbäumen wurde hier nicht eingegangen. Wie oben erwähnt, liegen in den Genotypisierungsdaten der LRS-Studie wohl auch keine der in Kapitel 4.3 beschriebenen Stammbaumfehler vor. Dies kann zum einen auf die gute Studienführung vor der Erhebung der Daten zurückzuführen sein, zum anderen auch durch das Studiendesign begründet werden: Es wurden nur Kernfamilien betrachtet. Fehlerquellen wie verzweigte Familienstrukturen, unbekannte Adoptionen oder ähnliches fallen dadurch weg.
Statistische Methoden, die als Ergänzung zur Mendel-Prüfungen zur Eingrenzung von Stammbaumfehlern geeignet sind, sind z. B. von McPeek und Sun beschrieben (McPeek & Sun, 2000; Sun et al., 2002).
Um auszuschließen, dass einer Person während der Genotypisierung ein falsches Geschlecht zugeordnet wurde, sollte von Hand überprüft werden, ob die Geschlechtsangaben in den Genotypisierungsdaten mit denen bei der Erhebung der Daten bestimmten Geschlechtsangaben übereinstimmen. Dieser Fehler kann unter Umständen während der Mendel-Checks und während der Verwendung von Testmethoden nach McPeek und Sun übersehen werden.

Abhängig von den Markern und den verwendeten Methoden, werden Fehler mit unterschiedlicher Häufigkeit entdeckt. Von der Rate der entdeckten Fehler kann die wahre Fehlerrate geschätzt werden.
Für SNP-Marker berichten Gordon et al. von entdeckten Fehlerraten zwischen 25% und 30% (Gordon et al., 1999), während sich bei Geller und Ziegler Raten von 39% bis 61% ergaben (Geller & Ziegler, 2002). Douglas et al. berichten von entdeckten Fehlerraten von 51% bis 77% für multi-allelische Marker und 13% bis 75% für di-allelische Marker (Douglas et al., 2002). Sie untersuchten drei Möglichkeiten, die zu Genotypisierungsfehlern führen können. Die erste Möglichkeit war der Fehler, dass ein homozygoter Genotyp als anders homozygot erfasst wird. Dieser Fehler kann als ein zufälliger Fehler angesehen, der nach den Häufigkeiten der Genotypen unter Annahme vom HWE auftreten kann. Die zweite Gruppe der untersuchten Fehler sind heterozygote Genotypen, die als falsch-homozygot angesehen

werden. Solche Fehler resultieren beispielsweise bei fehlerhafter PCR-Amplifikation. Als dritte Gruppe wurden die Fehler untersucht, dass ein homozygoter Genotyp fälschlich als falsch-heterozygot bezeichnet wird. Derartige Fehler können beim Auftreten von Stotterbanden in der PCR entstehen: Neben dem echten (homozygoten) Allel wird eine Stotterbande als falsches zweites Allel identifiziert.
Am häufigsten kann der Fehler der anders homozygoten Allele entdeckt werden, am seltensten der Fehler falsch-heterozygot statt homozygot. Dieser Fehler kann nur entdeckt werden, wenn beide Eltern für das Allel homozygot sind.
Generell gilt, dass die Wahrscheinlichkeit, einen Genotypisierungsfehler zu entdecken, von der Allelverteilung und Allelanzahl bei Eltern und Kindern abhängt. Je mehr Allele und je mehr genotypisierte Kinder zur Verfügung stehen, desto größer ist die Wahrscheinlichkeit, Fehler zu finden.

Auch wenn in praktisch allen Studien, die mit Genotypisierungsdaten arbeiten, Genotypisierungsfehler vorliegen, gibt es aber nur wenige Arbeiten, die sich dem Thema bisher angenommen haben (Hoffman & Amos, 2005; Pompanon et al., 2005).
Immer noch werden Studien durchgeführt, ohne dass eine Qualitätskontrolle stattfindet. Dabei ist erwiesen, das auch geringe Fehlerraten, große Auswirkungen auf die Informativität einer Untersuchung haben (Douglas et al., 2000; Akey et al., 2001).
Genotypisierungsfehler können zu falschen Schlussfolgerungen hinsichtlich der Allelfrequenz, der Kartierung, des Kopplungsungleichgewichts und der Markerabstände führen. Folgen sind verfälschte Ergebnisse und reduzierte Power bei Kopplungs- und Assoziationsstudien (Abecasis et al., 2001; Hao et al., 2004; Ziegler & König, 2006).
Für das Jahr 2003 werteten Bonin et al. alle Studien im Journal *Molecular Ecology* aus, in denen mit Genotypisierungsdaten gearbeitet wurde. Von 125 Veröffentlichungen, die mit Mikrosatelliten arbeiteten, wurde nur bei 6% eine Angabe über eine Fehlerrate oder Allelverlust bzw. falsche Allelamplifikation gemacht (Bonin et al., 2004).
Viele Autoren schlagen vor, standardmäßig alle Untersuchungen einer Kontrolle auf Genotypisierungsfehler zu unterziehen und entdeckte Fehlerraten in Veröffentlichungen anzugeben. Pompanon et al. propagieren dafür ein universelles Fehlermaß, die Fehlerrate pro Locus (Pompanon et al., 2005). Hiermit wären die Fehlerraten in verschiedenen Studien und auch zwischen verschiedenen Markern vergleichbar.

Finden Kontrollen auf Genotypisierungsfehler statt, gibt es prinzipiell zwei Möglichkeiten auf entdeckte Fehler zu reagieren. Die erste Möglichkeit ist der Ausschluss häufig betroffener Marker. In dieser Studie wurden die betroffenen Personen an den entsprechenden Loci als unbekannt („0 0“) genotypisiert. Diese Methode wird auch in den meisten anderen Studien angewandt (Sobel et al., 2002). Die Alternative wäre die betroffenen Marker für alle Familien/Personen erneut zu genotypisieren, wie es von Seaman und Holmans vorgeschlagen wird (Seaman & Holmans, 2005). Nachteil des Ausschlusses von Loci ist ein Verlust an Power, Nachteil der Regenotypisierung ist der erhöhte (finanzielle) Aufwand.

Der Bedarf für eine standardmäßige Qualitätssicherung besteht. Auch in Zukunft wird man sich mit dem Thema der Genotypisierungsfehler beschäftigen müssen. Die zunehmende Nutzung von SNPs in Untersuchungen verschärft das Problem der Genotypisierungsfehler. Bei der Nutzung von SNPs als genetische Marker fallen größere Datenmengen als bei Mikrosatelliten an. Und Genotypisierungsfehler können bei den SNPs als diallelische Marker viel eher Mendel-kompatibel sein. Sie wären also noch schwerer zu entdecken (Sobel et al., 2002).
Die optimale Strategie Fehler einzugrenzen, hängt immer vom Studientyp und von den benutzten Marker – Mikrosatelliten oder SNPs – ab (Pompanon et al., 2005). Zur Durchführung von Mendel-Checks und Kontrollen auf falsche Doppelrekombinanten etwa werden die Genotypen der Eltern benötigt.
Die in dieser Arbeit vorgestellten Methoden eignen sich so besonders für familienbasierte Untersuchungen, in denen Mikrosatelliten verwendet werden. Sie ermöglichen eine Reduktion der Fehlerquote in den Genotypisierungsraten. Die Datenqualität wird damit erhöht und die anschließenden Kopplungsanalysen ergeben validierte Ergebnisse.
Die hier vorgestellten Methoden können zusammen als ein Art „Leitfaden“ oder „Handbuch zur Qualitätskontrolle“ von Genotypisierungsdaten genutzt werden.

13 Zusammenfassung

Diese Arbeit entstand im Rahmen einer multizentrischen Studie, die nach möglichen neuen Genorten der Dyslexie sucht und schon beschriebene Loci bestätigen möchte (Remschmidt et al., 2000).

Die Dyslexie ist eine Erkrankung multifaktorieller Genese. Unter anderem liegt der Erkrankung eine genetische Komponente zugrunde (Schumacher et al., 2006).

Grundlage der Studie sind die Phänotypisierung und Genotypisierung von Probanden und ihren Geschwistern sowie die Genotypisierung ihrer Eltern. Mit Hilfe dieser Daten kann in einer Kopplungsanalyse nach genetischen Markern gesucht werden, die mit der Erkrankung gekoppelt auftreten und so auf Loci eines Dyslexie-verursachenden Gens schließen lassen.

In praktisch alle Studien, die sich mit Genotypisierungsdaten befassen, liegen Fehler in diesen Daten vor (Bonin et al., 2004; Pompanon et al., 2005). Diese Fehler können die Ergebnisse der Studien nachhaltig beeinträchtigen oder verfälschen (Douglas et al., 2000). Die Vermeidung und die Suche nach diesen Fehlern stellen eine schwierige Aufgabe dar. Von manchen Forschungsgruppen werden sie vernachlässigt (Pompanon et al., 2005), und in wenigen Studien finden sich Angaben darüber, ob Methoden zur Fehlerentdeckung verwendet wurden oder wie hoch die Rate an gefundenen Fehlern war (Bonin et al., 2004).

Zwei Arten von Fehlern lassen sich grundsätzlich unterscheiden: Stammbaumfehler und Genotypisierungsfehler. In dieser Arbeit geht es vor allem um die Genotypisierungsfehler. Es werden verschiedene Methoden betrachtet, mit denen Genotypisierungsfehler aufgespürt werden sollen, um so die Datenqualität zu erhöhen.

Am Beispiel der LRS-Studie sind auf die Genotypisierungsdaten verschiedene Testmethoden angewandt worden. Wichtige Ansätze waren die Nutzung von Mendel-Checks (O'Connell & Weeks, 1998), die Suche nach Doppelrekombinanten (Ziegler & König, 2006), Kontrollen während der Kartierung, zwei verschiedene Methoden, Abweichungen von der Hardy-Weinberg Verteilung in den Daten zu entdecken, und die Suche nach Nullallelen.

Mit den vorgestellten Methoden zur Qualitätssicherung kann die Fehlerquote in Genotypisierungsraten minimiert und somit die Datenqualität erhöht werden. Dieses führt zu validierten Ergebnissen in den anschließenden Kopplungsanalysen. Mit der Zusammenstellung geeigneter Maßnahmen zur Qualitätssicherung bildet diese Arbeit damit die Grundlage für eine Standardarbeitsanweisung (engl. *Standard operating procedure*, SOP) für die Qualitätssicherung bei familienbasierten Studien mit Mikrosatelliten.

14 Literaturverzeichnis

Abecasis, G.R., Cherny, S.S., Cardon, L.R. (2001). “The impact of genotyping error on family-based analysis of quantitative traits.” *Eur J Hum Genet*, 9: 130-134.

Abu-Rabia, S. & Maroun, L. (2005). "The effect of consanguineous marriage on reading disability in the Arab community." *Dyslexia*, 11: 1-21.

Akey, J.M., Zhang, K., Xiong, M., Doris, P., Jin, L. (2001). "The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures." *Am J Hum Genet*, 68: 1447-1456.

Allison, D.B., Heo, M., Schork, N.J., Wong, S.L., Elston, R.C. (1998). "Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power." *Hum Hered*, 48: 97-107.

Angier, N. (1997). “Sex and the female chimp.” *The New York Times*, 27. Mai: C8.

Angier, N. (2001). “A fresh look at the straying ways of the female chimp.” *The New York Times*, 15. Mai: F3.

A.P.A., American Psychiatric Association (1994). “Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition (DSM-IV)”. American Psychiatric Association, Washington D.C.

Bellis, M.A., Hughes, K., Hughes, S. & Ashton, J.R. (2005). "Measuring paternal discrepancy and its public health consequences." *J Epidemiol Community Health*, 59: 749-754.

Bonin, A., Bellemain, E., et al. (2004). "How to track and assess genotyping errors in population genetics studies." *Mol Ecol*, 13: 3261-3273.

Botstein, D., White, R.L., et al. (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms." *Am J Hum Genet*, 32: 314-331.

Broman, K.W. et al. (1998). “Comprehensive human genetic maps: individual and sex-specific variation in recombination.” *Am J Hum Genet*, 63: 861-869.

Brookfield, J. F. (1996). "A simple new method for estimating null allele frequency from heterozygote deficiency." *Mol Ecol*, 5: 453-5.

Butler, J. M. (2001). "Forensic DNA typing: biology and technology behind STR markers." Academic Press, San Diego.

Cardon, L.R., Smith, S.D., Fulker, D.W., Kimberling, W.J., Pennington, B.F., DeFries, J.C. (1994). "Quantitative trait locus for reading disability on chromosome 6." *Science*, 266: 276-279.

CHLC - The Cooperative Human Linkage Center (1999). World Wide Web URL:

http://lpgws.nci.nih.gov/CHLC/

Cichon, S., Freudenberg, J., Propping, P., Nöthen, M.M. (2002). „Variabilität im menschlichen Genom – Bedeutung für die Krankheitsforschung.“ *Deutsches Ärzteblatt*, 99: A3091-3101.

Chakraborty, R., De Andrade, M., et al. (1992). "Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications." *Ann Hum Genet*, 56: 45-57.

Cope, N., Harold, D., Hill, G., Moskvina, V., Stevenson, J., Holmans, P., Owen, M.J., O'Donovan, M., Williams, J. (2005). "Strong evidence that KIAA0319 on chromosome 6p is a susceptibility gene for developmental dyslexia." *Am J Hum Genet*, 76: 581–591.

Deffenbacher, K.E., Kenyon, J.B., Hoover, D.M., Olson, R.K., Pennington, B.F., DeFries, J.C., Smith, S.D. (2004). "Refinement of the 6p21.3 quantitative trait locus influencing dyslexia: linkage and association analyses." *Hum Genet*, 115: 128-138.

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

Douglas, J. A., Boehnke, M. et al. (2000). "A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data." *Am J Hum Genet*, 66: 1287-1297.

Douglas, J. A., Skol, A. D., et al. (2002). "Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data." *Am J Hum Genet*, 70: 487-495.

Downfreq (1995). Version 1.1. Weeks, D., Division of Statistical Genetics, Department of Human Genetics, University of Pittsburgh. World Wide Web URL: http://watson.hgen.pitt.edu/docs/Downfreq2-Document.html

Esser, G., Wyschkon, A., et al. (2002). "Was wird aus Achtjährigen mit einer Lese- und Rechtschreibstörung - Ergebnisse im Alter von 25 Jahren." *Z Klin Psychol Psychiat Psychother*: 235-242.

Fagerheim, T., et al. (1999). "A new gene (DYX3) for dyslexia is located on chromosome 2." *J Med Genet*, 36: 664-669.

Field, L.L., Kaplan, B.J. (1998). "Absence of linkage of phonological coding dyslexia to chromosome 6p23-p21.3 in a large family data set." *Am J Hum Genet*, 63: 1448-1456.

Fisher, S. A., Lewis, C. M., et al. (2001). "Detecting population outliers and null alleles in linkage data: application to GAW12 asthma studies." *Genet Epidemiol*, 21: 18-23.

Fisher, S.E., Marlow, A.J., Lamb, J., Maestrini, E., Williams, D.F., Richardson, A.J., Weeks, D.E., Stein, J.F., Monaco, A.P. (1999). "A quantitative-trait locus on chromosome 6p influences different aspects of developmental dyslexia." *Am J Hum Genet*, 64: 146-156.

Fisher, S.E., Francks, C., Marlow, A.J., MacPhie, I.L., Newbury, D.F., Cardon, L.R., Ishikawa-Brush, Y., Richardson, A.J., Talcott, J.B., Gayan, J., Olson, R.K., Pennington, B.F., Smith, S.D., DeFries, J.C., Stein, J.F., Monaco, A.P. (2002). "Independent genome-wide scans identify a chromosome 18 quantitative-trait locus influencing dyslexia." *Nat Genet*, 30: 86-91.

Francks, C., MacPhie, I. L., et al. (2002). "The genetic basis of dyslexia." *Lancet Neurol*, 1: 483-490.

Francks, C., Paracchini, S., Smith, S.D., Richardson, A.J., Scerri, T.S., Cardon, L.R., Marlow, A.J., MacPhie, I.L., Walter, J., Pennington, B.F., Fisher, S.E., Olson, R.K., DeFries, J.C., Stein, J.F., Monaco, A.P. (2004). "A 77-kilobase region of chromosome 6p22.2 is associated with dyslexia in families from the United Kingdom and from the United States." *Am J Hum Genet*, 75: 1046-1058.

Gagneux, P., Woodruff, D.S. & Boesch, C. (1997). „Furtive mating in female chimpanzees.“ *Nature*, 387: 358 – 359.

Gayan, J., Smith, S.D., Cherny, S.S., Cardon, L.R., Fulker, D.W., Brower, A.M., Olson, R.K., Pennington, B.F., DeFries, J.C. (1999). "Quantitative-trait locus for specific language and reading deficits on chromosome 6p." *Am J Hum Genet*, 64: 157-164.

GDB – The Human Genome Database. World Wide Web URL: http://gdbwww.gdb.org/

Geller, F. & Ziegler, A. (2000). "Studiendesigns zur Rekrutierung von Geschwisterpaaren für die genetische Kartierung quantitativer Phänotypen." *Med Genet*, 12: 423 - 427.

Geller, F. & Ziegler, A. (2002). "Detection rates for genotyping errors in SNPs using the trio design." *Hum Hered*, 54: 111-117.

Genehunter (2006). Version 2.1 r6. Kruglyag, L., Fred Hutchison Cancer Research Center. Worl Wide Web URL: http://www.fhcrc.org/science/labs/kruglyak/Downloads/index.html

Ghebranious, N., Vaske, D., et al. (2003). "STRP screening sets for the human genome at 5 cM density." *BMC Genom*ics, 4: 6.

Gilger, J.W., Borecki, I.B., DeFries, J.C., Pennington, B.F. (1994). "Commingling and segregation analysis of reading performance in families of normal reading probands." *Behav Genet*, 24:345-55.

Gordon, D., Heath, S. C., et al. (1999). "True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms." *Hum Hered*, 49: 65-70.

Green, P., Falls, K. und Crook, S. (1990). "Documentation for CRI-MAP, version 2.4". Washington University School of Medicine, St. Louis, Missouri.

Grigorenko, E.L. (1997). "Susceptibility loci for distinct components of developmental dyslexia on chromosomes 6 and 15." *Am J Hum Genet*, 60: 27-39.

Grigorenko, E.L., Wood, F.B., Meyer, M.S., Pauls, J.E., Hart, L.A., Pauls, D.L. (2001). "Linkage studies suggest a possible locus for developmental dyslexia on chromosome 1p." *Am J Med Genet*, 105: 120-129.

Grigorenko, E.L., Wood, F.B., Golovyan, L., Meyer, M., Romano, C., Pauls, D. (2003). "Continuing the search for dyslexia genes on 6p." *Am J Med Genet B Neuropsychiatr Genet*, 118: 89-98.

Guo, S. W. & Thompson, E.A. (1992). "Performing the exact test of Hardy-Weinberg proportion for multiple alleles." *Biometrics*, 48: 361-372.

Hao K., Li C., Rosenow C., Hung Wong W. (2004). “Estimation of genotype error rate using samples with pedigree information--an application on the GeneChip Mapping 10K array.” *Genomics*, 84: 623-630.

Hennig, W. (2002). “Genetik.“ 4. Auflage, Springer-Verlag, Heidelberg.

Hsiung, G.Y., Kaplan, B.J., Petryshen, T.L., Lu, S., Field, L.L. (2004). "A dyslexia susceptibility locus (DYX7) linked to dopamine D4 receptor (DRD4) region on chromosome 11p15.5." *Am J Med Genet B Neuropsychiatr Genet*, 125: 112-119.

Hoffman, J.I. & Amos, W. (2005). "Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion." *Mol Ecol*, 14:599-612.

Igo, R.P. Jr., Chapman, N.H., Berninger, V.W., Matsushita, M., Brkanac, Z., Rothstein, J.H., Holzman, T., Nielsen, K., Raskind, W.H., Wijsman, E.M. (2006). "Genomewide scan for real-word reading subphenotypes of dyslexia: novel chromosome 13 locus and genetic complexity." *Am J Med Genet B Neuropsychiatr Genet*, 141:15-27.

Kaminen, N., Hannula-Jouppi, K., Kestila, M., Lahermo, P., Muller, K., Kaaranen, M., Myllyluoma, B., Voutilainen, A., Lyytinen, H., Nopola-Hemmi, J., Kere, J. (2003). "A genome scan for developmental dyslexia confirms linkage to chromosome 2p11 and suggests a new locus on 7q32." *J Med Genet*, 40: 340-345.

Kaplan, D.E., Gayan, J., Ahn, J., Won, T.W., Pauls, D., Olson, R.K., DeFries, J.C., Wood, F., Pennington, B.F., Page, G.P., Smith, S.D., Gruen, J.R. (2002). "Evidence for linkage and association with reading disability on 6p21.3-22." *Am J Hum Genet*, 70: 1287-1298.

Katusic, S.K., Colligan, R.C., Barbaresi, W.J., Schaid, D.J., Jacobsen, S.J. (2001). "Incidence of reading-disability in a population-based birth cohort, 1976 -1982." *Mayo Clin Proc*, 76: 1081-1092.

Kong, A., et al. (2002). "A high-resolution recombination map of the human genome." *Nat Genet*, 31: 225-6.

de Kovel, C.G., Hol, F.A., Heister, J.G., Willemen, J.J., Sandkuijl, L.A., Franke, B., Padberg, G.W. (2004). "Genomewide scan identifies susceptibility locus for dyslexia on Xq27 in an extended Dutch family." *J Med Genet*, 41: 652-657.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., Lander, E.S. (1996). "Parametric and nonparametric linkage analysis: a unified multipoint approach". *Am J Human Genetic*, 58:1347-1363.

Lee, W.C. (2003). "Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals." *Am J Epidemiol*, 158: 397-400.

Lewitter, F.I., DeFries, J.C., et al. (1980). "Genetic models of reading disability." *Behav Genet*, 10: 9-30.

Manaster, C.J., Nanthakumar, E., Morin, P.A. (1999). „Detecting null alleles with Vasarely charts." Proceedings of the 10th IEEE Visualization 1999 Conference (VIS '99), IEEE Computer Society.

Marshfield Clinic (2005). Center for Medical Genetics. World Wide Web URL: http://research.marshfieldclinic.org/genetics/

Marx, H., Jansen, H., Mannhaupt, G., Skowronek, H., Näslund, J.C., Schneider, W. (1993). "Prediction of difficulties in reading and spelling on the basis of the Bielefeld screening." In Grimm, H., Skowronek, H. (Hrsg.): "Language acquisition problems and reading disorders aspects of diagnosis and intervention." Walter de Gruyter, Berlin.

McPeek, M.S. & Sun, L. (2000). "Statistical tests for detection of misspecified relationships by use of genome-screen data." *Am J Hum Genet*, 66: 1076-1094.

Miller, C. R., Joyce, P., et al. (2002). "Assessing allelic dropout and genotype reliability using maximum likelihood." *Genetics*, 160: 357-66.

Morris, D.W., Robinson, L., Turic, D., Duke, M., Webb, V., Milham, C., Hopkin, E., Pound, K., Fernando, S., Easton, M., Hamshere, M., Williams, N., McGuffin, P., Stevenson, J., Krawczak, M., Owen, M.J., O'Donovan, M.C., Williams, J. (2000). "Family-based association mapping provides evidence for a gene for reading disability on chromosome 15q." *Hum Mol Genet*, 9: 843-848.

Nakamura, Y., Leppert, M., et al. (1987). "Variable number of tandem repeat (VNTR) markers for human gene mapping." *Science*, 235: 1616-1622.

Neale, M.C., Neale, B.M., et al. (2002). "Nonpaternity in linkage studies of extremely discordant sib pairs." *Am J Hum Genet*, 70: 526-9.

Nopola-Hemmi, J., Myllyluoma, B., Haltia, T., Taipale, M., Ollikainen, V., Ahonen, T., Voutilainen, A., Kere, J., Widen, E. (2001). "A dominant gene for developmental dyslexia on chromosome 3." *J Med Genet*, 38: 658-664.

NullAlleleCheck (2004). Version 0.98. Franke, D., Heenes, V., Kleensang, A. Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein – Campus Lübeck, Lübeck.

O'Connell, J.R. & Weeks, D.E. (1998). "PedCheck: a program for identification of genotype incompatibilities in linkage analysis." *Am J Hum Genet*, 63: 259-66.

O'Connell, J.R. & Weeks, D.E. (1999). "An optimal algorithm for automatic genotype elimination." *Am J Hum Genet*, 65: 1733-40.

Olson, R.K., Forsberg, H., Wise B. (1994): “Genes, environment, and development of orthographic skills.” In Berninger, V.W. (Hrsg.): “The varieties of orthographic knowledge I: theoretical and developmental issues.” Kluwer, Dordrecht: 27-71.

OMIM Database: Online Mendelian Inheritance in Man. World Wide Web URL: http://www3.ncbi.nlm.nih.gov/

Van Oosterhout, C., et al. (2004). "Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data." *Mol Ecol*, 4: 535-538.

Ott, J. (1999). "Analysis of human genetic linkage." 3. Auflage, Johns Hopkins University Press, Baltimore.

Pedcheck (2000). Version 1.0. O’Connell, J. Division of Statistical Genetics, Department of Human Genetics, University of Pittsburgh. World Wide Web URL: http://watson.hgen.pitt.edu/register/.

Pedstats, Version 0.6.4, Center for Statistical Genetics, University of Michigan. World Wide Web URL: http://www.sph.umich.edu/csg/abecasis/PedStats/download/

Pennington, B. F., Gilger, J. W., et al. (1991). "Evidence for major gene transmission of developmental dyslexia." *JAMA*, 266: 1527-1534.

Perlin, M. W., Lancia, G., Ng, S. K. (1995). "Toward fully automated genotyping: genotyping microsatellite markers by deconvolution." *Am J Hum Genet*, 57: 1199–1210.

Petryshen, T.L., Kaplan, B.J., Liu, M.F., Field, L.L. (2000). "Absence of significant linkage between phonological coding dyslexia and chromosome 6p23-21.3, as determined by use of quantitative-trait methods: confirmation of qualitative analyses." *Am J Hum Genet*, 66:708-714.

Petryshen, T.L., Kaplan, B.J., Fu Liu, M., de French, N.S., Tobias, R., Hughes, M.L., Field, L.L. (2001). "Evidence for a susceptibility locus on chromosome 6q influencing phonological coding dyslexia." *Am J Med Genet*, 105: 507-517.

Pompanon, F., Bonin, A., Bellemain, E., Taberlet, P. (2005). „Genotyping errors: causes, consequences and solutions.“ *Nature Reviews Genetic*, 6: 847-846.

Pringle-Morgan, W. (1896). "A case of congenital word blindness." Br J Med, 2: 1378.

R (2005). Version 2.2.1. The R Foundation for Statistical Computing.

Rabbata, S. & Richter-Kuhlmann, E.A. (2005). "Unethisch und bedenklich." *Deutsches Ärzteblatt*, 102: A89-A90.

Rabin, M., Wen, X.L., Hepburn, M., Lubs, H.A., Feldman, E., Duara, R. (1993). "Suggestive linkage of developmental dyslexia to chromosome 1p34-p36." *Lancet*, 342: 178.

Remschmidt, H., et al. (2000). „Neurobiologische und molekulargenetische Untersuchungen zur Lese-Rechtschreibstörung (Multizentrisches Forschungsprojekt).“ Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck.

Remschmidt, H. (2002). "Hyperkinetisches Syndrom und Legasthenie: Diagnostik, Ätiologie, Krankheitsverlauf und Behandlung." *Deutsches Ärzteblatt*, 99: A-2632/ B-2242/ C-2105.

Ridley, M. (2003). „Evolution.” Blackwell Science Inc.

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

Rutter, M. (2004) „Sex differences in developmental reading disability: new findings from 4 epidemiological studies.” *JAMA*, 291: 2007-2012.

Schumacher, J., König, I.R., Plume, E., Propping, P., Warnke, A., Manthey, M., Duell, M., Kleensang, A., Repsilber, D., Preis, M., Remschmidt, H., Ziegler, A., Nothen, M.M.,

Schulte-Körne, G. (2005). "Linkage analyses of chromosomal region 18p11-q12 in dyslexia." J Neural Transm, 113: 417-423.

Schumacher, J., Anthoni, H., Dahdouh, F., König, I.R., Hillmer, A.M., Kluck, N., Manthey, M., Plume, E., Warnke, A., Remschmidt, H., Hulsmann, J., Cichon, S., Lindgren, C.M., Propping, P., Zucchelli, M., Ziegler, A., Peyrard-Janvid, M., Schulte-Körne, G., Nöthen, M.M., Kere, J. (2006). "Strong Genetic Evidence of DCDC2 as a Susceptibility Gene for Dyslexia." *Am J Hum Genet*, 78: 52-62.

Schulte-Körne, G. (1998a). "Zur Genetik der Lese-Rechtschreibstörung (Legasthenie)." *Med Genet*, 10: 402-405.

Schulte-Körne, G. (1998b). "Evidence for linkage of spelling disability to chromosome 15." *Am J Hum Genet*: 279-282.

Schulte-Körne, G. (2001a). "Annotation: Genetics of reading and spelling disorder." *J Child Psychol Psychiatry*, 42: 985-997.

Schulte-Körne, G. (2001b). "Genetic aspects of dyslexia." *J Child Psychol Psychiatry*, 42: 163-173.

Schulte-Körne, G. (2001c). „Lese-Rechtschreibschwäche und Sprachwahrnehmung.“ Waxmann, Münster.

Schulte-Körne, G. & Remschmidt, H. (2003). "Legasthenie - Symptomatik, Diagnostik, Ursachen, Verlauf und Behandlung." *Deutsches Ärzteblatt*, 100: 396-406.

Seaman, S.R. & Holmans, P. (2005). "Effect of genotyping error on type-I error rate of affected sib pair studies with genotyped parents." *Hum Hered*, 59: 157-164.

Seyffert, W. (1998). „Lehrbuch der Genetik.“ Gustav Fischer Verlag, Stuttgart.

Shaywitz, S.E., Shaywitz, B.A., et al. (1990). "Prevalence of reading disability in boys and girls. Results of the Connecticut Longitudinal Study." *JAMA*, 264: 998-1002.

Shinde, D., Lai, Y., Sun, F., Arnheim, N. (2003). "Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites." *Nucleic Acids Res*, 31: 974–980.

Smith, S.D. (1983). "Specific reading disability; identification of an inherited form through linkage analysis." *Science*, 219: 1345-1347.

Smith, S.D. (1991). "Screening for multiple genes influencing dyslexia." *Read Writ Interdisc J*, 3: 285-298.

Sobel, E., Papp, J.C., Lange, K. (2002). "Detection and integration of genotyping errors in statistical genetics." *Am J Hum Genet*, 70: 496-508.

Stevenson, J., Fredman, G. (1990). "The social environmental correlates of readind ability." J *Child Psychol Psychiatry*, 31: 681-698.

Stevenson, J. (1991). "Which aspects of processing text mediate genetic effects?" *Read Writ Interdisc J*, 3: 249-269.

Strachan, T. & Read A.P. (1999). “Human Molecular Genetics 2.” John Wiley & Sons Inc., New York.

Strachan, T. & Read A.P. (2004). “Human Molecular Genetics 3.” Garland Science, New York.

Strehlow, U., Kluge, R., Möller, H. & Haffner, J. (1992). "Der langfristige Verlauf der Legasthenie über die Schulzeit hinaus: Katamnesen aus einer Kinderpsychiatrischen Ambulanz." *Z Kinder Jugendpsychiatr*, 20: 254-265.

Sun, L., Wilder, K., McPeek, M.S. (2002). “Enhanced pedigree error detection.” *Hum Hered*, 54; 99-110.

Tautz, D. & Renz, M. (1984). "Simple sequences are ubiquitous repetitive components of eukaryotic genomes." *Nucleic Acids Res*, 12: 4127-4138.

Terwilliger, J.D. & Ott, J. (1994) "Handbook of Human Genetic Linkage" Johns Hopkins University Press, Baltimore.

Torgesen, J. K., Morgan, S. & Davis, C. (1992). "Effects of two types of phonological awareness training on word learning in kindergarten children." *J Educ Psychol*, 84: 364-370.

Tzenova, J., Kaplan, B.J., Petryshen, T.L., Field, L.L. (2004). "Confirmation of a dyslexia susceptibility locus on chromosome 1p34-p36 in a set of 100 Canadian families." *Am J Med Genet B Neuropsychiatr Genet*, 127: 117-124.

UCSC Genome Browser. Genome Bioinformatics Group of UC Santa Cruz. World Wide Web URL:
http://genome.ucsc.edu/cgi-bin/hgGateway

Wang, D.G., Fan, J.B., et al. (1998). "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome." *Science*, 280: 1077-1082.

Warnke, A. & Roth, E. (2000). „Umschriebene Lese-Rechtschreibstörung. Lehrbuch der klinischen Kinderpsychologie.“ 4. Auflage, Hogrefe, Götingen: 453-476.

Weber, J.L. & Broman, K.W. (2001). "Genotyping for human whole-genome scans: past, present, and future." *Adv Genet*, 42: 77-96.

Wellek, S. (2004). "Tests for establishing compatibility of an observed genotype distribution with Hardy-Weinberg equilibrium in the case of a biallelic locus." *Biometrics*, 60: 694-703.

WHO (1992). "International Statistical Classification of Diseases and Related Health Problems." 10. Neubearbeitung, Band 1, WHO, Genf.

Wigginton, J.E.& Abecasis, G.R. (2005). "Pedstats: descriptive statistics, graphics ang quality assessment for gene mapping data." *Bioinformatics*, 21:3445-3447.

Willcutt, E.G., Pennington, B.F., et al. (2002). "Quantitative trait locus for reading disability on chromosome 6p is pleiotropic for attention-deficit/hyperactivity disorder." *Am J Med Genet*, 114: 260-268.

Ziegler, A. (1999). "Sampling strategies for model free linkage analyses of quantitative traits: implications for sib pair studies of reading and spelling disabilities to minimize the total study cost." *Eur Child Adolesc Psychiatry*, 8: 35-39.

Ziegler, A., König, I.R., Deimel, W., Plume, E., Nöthen, M.M., 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.

Ziegler, A. & König, I.R. (2006). „A Statistical Approach to Genetic Epidemiology: Concepts and Applications." Wiley-VCH, New York.

Ziegler, A., Barth, N., Coners, H., Mayer, H., Hebebrand, J. (2006). "Practical considerations on the use of extreme sib-pairs for obesity." *Methods Inf Med*, 46: in press.

15 Anhang

15.1 Danksagung

Dem Direktor des Institutes, Herrn Prof. Dr. rer. nat. Andreas Ziegler, danke ich sehr für die Bereitstellung des Themas und damit die Möglichkeit, dass ich diese Dissertation verfassen konnte. Ihm habe ich für die freundliche und fachliche Unterstützung bei der Ausarbeitung dieser Arbeit zu danken.

Ich möchte mich bei Frau Dr. rer. biol. hum. Inke R. König für ihre intensive Betreuung während der gesamten Entstehung dieser Arbeit sehr herzlich bedanken. Durch sie erfuhr ich eine gute Einarbeitung in die Thematik, sie nahm sich stets Zeit, und ich verdanke ihr viele Verbesserungsvorschläge.

Allen Mitarbeitern des Institutes für Medizinische Biometrie und Statistik möchte ich für ihre freundliche Aufnahme danken. Für ihre Unterstützung bei der Umsetzung dieser Arbeit möchte ich besonders Herrn Dipl.-Biochem. André Kleensang und Herrn Dipl.-Inf. Friedrich Pahlke danken. Für ihre Korrekturen danke ich Frau Dipl.-Stat. (FH) Anika Götz.

Meiner Familie und insbesondere meiner Schwester Annette und ihrem Mann, Dr. med. Björn Harder, möchte ich für ihre mentale Unterstützung und die wertvollen Hinweise danken, die zum Gelingen dieser Arbeit beitrugen. Meinem Vater danke ich sehr dafür, dass er mir mein Studium ermöglicht hat.

Außerdem möchte ich allen weiteren Personen danken, die mich bei der Fertigstellung dieser Arbeit auf vielfältige Weise unterstützt und motiviert haben.

15.2 Lebenslauf

Name	Kiewert
Vorname	Alexander
Geburtstag	11. Juli 1978
Geburtsort	Osnabrück

Studium

seit Oktober 2000	Studium der Medizin in Lübeck, Ärztliche Vorprüfung im April 2003
Oktober 1999 - August 2000	zwei Semester Maschinenbaustudium an der TU München und 3-monatiges Industriepraktikum

Dienstzeit

Juli 1998 - April 1999	10-monatiger Grundwehrdienst in Lingen(Ems)

Schulbesuch

1991 – 98	Gymnasium Carolinum Osnabrück, Abitur 1998
1989 – 91	Orientierungsstufe Dom, Osnabrück
1985 – 89	Erich-Kästner-Schule, Wallenhorst

Famulaturen und Praktika	Unfallchirurgie, Innere Medizin, Radiologie, Dermatologie und Pathologie; Notarzteinsatzfahrzeug-Praktikum der Universität Lübeck; Auslandsfamulaturen in Spanien und Chile

Zeitlicher Rahmen der Dissertation

November 2003	Vergabe des Themas und Literaturrecherche
April 2004	Beginn der praktischen Bearbeitung
Juli 2005	Beginn der schriftlichen Ausarbeitung
September 2006	Abgabe der Arbeit