Statistische Methoden zur familienbasierten Assoziationsanalyse

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Andreas Hahn

aus

Lüdenscheid

Bonn 2000

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Prof. Dr. Max P. Baur
2. Referent: Prof. Dr. Wolfgang Alt
Tag der Promotion: 15.8.2000

Meinen Eltern

Inhaltsverzeichnis

Kapitel 1

Einleitung

Für die Entwicklung von Medikamenten ist es wichtig zu verstehen, wie im Körper eines gesunden Menschen die verschiedenen Abläufe funktionieren und zusammenwirken. Einfluss auf die Abläufe haben sowohl genetische Faktoren als auch die Umwelt. Ein Schlüssel zum Verständnis der Zusammenhänge sind die Gene, die das Risiko für eine bestimmte Krankheit beim Menschen erhöhen oder verringern. Bei Krankheiten, die durch genau ein Gen bestimmt werden, lässt sich in vielen Fällen mittels Methoden der mathematischen Genetik feststellen, auf welchem der 23 menschlichen Chromosomen dieses Gen liegt. Mit diesen Methoden ist dann auch eine noch genauere Lokalisierung des Gens auf eine Region des Chromosoms möglich. Durch Genotypisierung erhält man die DNA-Basenabfolge sowohl für Patienten als auch für Kontrollpersonen. Die Unterschiede in den Basenabfolgen und die dadurch verursachten Unterschiede in den Proteinsequenzen können beispielsweise dazu führen, dass die Menge eines für einen Stoffwechselweg notwendigen Enzyms in den Patienten erhöht oder erniedrigt ist, was Krankheitssymptome nach sich ziehen kann. Durch entsprechende medizinische Maßnahmen (Einnahme von Medikamenten oder Durchführung einer Gentherapie) wird es dann eventuell möglich sein, die Krankheitssymptome in den Patienten zu mildern oder sogar völlig zu beseitigen. Auch für gesunde Menschen kann die Kartierung und Identifizierung krankheitsrelevanter Gene von Bedeutung sein. So können nach Erkennen eines Gens, welches das Risiko für eine Krankheit erhöht, präventive Maßnahmen

(z.B. die Ernährung oder den Bewegungsapparat betreffend) ergriffen werden. Bei der Stoffwechselkrankheit Phenylketonurie (Collins, 1999) beispielsweise, die unerkannt zu Schwachsinn führt, ist durch den Verzicht auf Phenylalanin eine normale Entwicklung gewährleistet.

Die Kartierung krankheitsrelevanter Gene erfolgt durch Kopplungs- und Assoziationanalyse. Diese beiden Paradigmen der mathematischen Genetik unterscheiden sich in folgenden Punkten: Mit der *Kopplungsanalyse* prüft man innerhalb von Familienstammbäumen auf gemeinsame Transmission von Allelen eines genetischen Markers (das ist ein Genort mit bekannter Position auf dem Chromosom) und der interessierenden Krankheit. Bei der *Assoziationsanalyse* prüft man innerhalb einer Population den Zusammenhang (allelische Assoziation) zwischen bestimmten Allelen am Marker und dieser Krankheit. Des weiteren ist die Kopplunganalyse eher für eine Grobkartierung des Krankheitsgenortes geeignet, während die Assoziationsanalyse eine genauere Positionierung zulässt (Feinkartierung). Schließlich kann man mit der Assoziationsanalyse, aufgrund eines geringeren notwendigen Stichprobenumfanges, besser als mit der Kopplungsanalyse Genorte identifizieren, die an der Entstehung komplexer Krankheiten beteiligt sind (Risch und Merikangas, 1996).

Markersysteme werden also sowohl bei der Kopplungsanalyse, als auch bei der Assoziationsanalyse verwendet. Wünschenswerte Eigenschaften von Markern sind eine niedrige Mutationsrate (da sie dann gute Orientierung bieten), eine hohe Dichte (weil die Bereiche ohne Information dann klein sind) und ein hoher Heterozygotiegrad (da nur heterozygote Eltern bezüglich Kopplung und Assoziation informativ sind). Außerdem sollten sie einfach, schnell und preisgünstig zu typisieren sein. Derzeit gibt es keinen Marker, der alle diese Eigenschaften besitzt.

Noch vor wenigen Jahren waren Mikrosatelliten-Marker die Marker, welche in der Regel zum Auffinden von Assoziation verwendet wurden. Mikrosatelliten-Marker haben einen durschnittlichen Abstand von 10^7 Basenpaaren und sind hochpolymorph (mehr als zehn Allele). Seit etwa 1997 werden sie mehr und mehr von biallelischen Markern, den sogenannten SNPs ("single nucleotide polymorphisms"), abgelöst. Diese haben einen sehr viel geringeren Abstand zueinander (auf etwa 500 Basenpaare kommt ein SNP). Weil die SNPs jeweils nur zwei Allele haben, sind mehrere (ungefähr fünf) SNPs notwendig, um die Information eines Mikrosatelliten-Markers zu ersetzen. Neben der hohen Dichte haben SNPs die Vorteile, dass sie eine geringe Mutationsrate aufweisen und sich schnell und automatisiert typisieren lassen.

In dieser Arbeit werden der Transmission/Disequilibrium-Test (TDT) und seine Erweiterungen beschrieben. Bei dem im zweiten Kapitel beschriebenen TDT handelt es sich um eine familienbasierte Methode zur Auffindung von Krankheitsgenorten mit Hilfe von Trios. Als "Trio" bezeichnet man eine Familie bestehend aus einem erkrankten Kind und seinen Eltern. Die Erweiterungen des TDT betreffen die Anzahl der Allele am Marker (multiallelischer TDT) und das "Ersetzen" oder die Rekonstruktion nicht typisierter, elterlicher Genotypen. In Kapitel drei wird die familienbasierte Assoziationsanalyse bei mehr als einem Marker beschrieben. Nach Verdeutlichung der Problemstellung werden die beiden Programme GENEHUNTER und Transmit vorgestellt, die eine entsprechende Analyse durchführen können. Abschließend wird ein neues Modell entwickelt, mit dem sich Assoziation zwischen einer Krankheit und einem Multi-Marker-Haplotyp testen lässt. Als "Haplotyp" bezeichnet man das Allelmuster für benachbarte Genorte auf dem gleichen Chromosom. Ziel dieses Testes ist das Auffinden von Genorten, deren Allele oder Haplotypen mit einer Krankheit assoziiert sind. Insbesondere können mit diesem Modell Assoziationsanalysen durchgeführt

werden, auch wenn die Übertragung von Haplotypen der Eltern auf die Kinder nicht eindeutig ist (unbekannte Phase). Als Anwendung wird sowohl ein simulierter als auch ein realer Datensatz analysiert. Dort zeigt sich, dass der Multi-Marker-TDT bei Vorliegen biallelischer Marker gegenüber Methoden, die nur *einen* Marker verwenden, vorzuziehen ist. Deshalb und weil die Typisierung von SNPs in den letzten Jahren stark zugenommen hat werden Multi-Marker-Methoden in Zukunft mehr und mehr an Bedeutung gewinnen.

Kapitel 2

Familienbasierte Assoziationsanalyse bei *einem* Marker

2.1 Problemstellung und Übersicht

Das wesentliche Ziel von Assoziationsstudien besteht darin, krankheits(mit)verursachende Gene im menschlichen Genom zu lokalisieren. Dazu verwendet man im Unterschied zu den Kopplungsstudien Markerdaten von Fällen und Kontrollen aus der Bevölkerung. Auch die Art des untersuchten Zusammenhanges weicht von der in Kopplungsstudien ab: Dort interessiert man sich eben für die Kopplung zwischen Marker und Krankheitsgenort und somit die gemeinsame Vererbung der Allele eines Markers mit einer Krankheit, während man in einer Assoziationsstudie bei Signifikanz einen Zusammenhang zwischen *bestimmten* Allelen und einer Krankheit nachgewiesen hat.

Definition: Allelische Assoziation
Mit dem Begriff "Allelische Assoziation" wird der Zusammenhang zwischen den Allelen zweier Genorte beschrieben. Zur Auffindung von Krankheitsgenorten interessiert man sich häufig für allelische Assoziation zwischen den Allelen eines Markers und eines Krankheitsgenortes. Diese wird unter Benutzung der folgenden Notation definiert.
Die relative Häufigkeit von Allel A_1 am Marker betrage a_1 und die relative Häufigkeit von Allel Q_1 am Krankheitsgenort betrage q_1. Das Ausmaß

allelischer Assoziation zwischen Allel A_1 am Marker und Q_1 am Krankheitsgenort wird mit δ bezeichnet und ist definiert als:

$$\delta := \text{Häufigkeit}(A_1, Q_1) - a_1 \cdot q_1$$

Sind die Allele der beiden Genorte unabhängig voneinander, so ist die relative Häufigkeit des Haplotyps A_1Q_1 gleich $a_1 \cdot q_1$. Das heißt dann, dass δ gleich Null ist und somit keine allelische Assoziation zwischen dem Allel A_1 und dem Allel Q_1 besteht. Falls der Haplotyp A_1Q_1 häufiger (seltener) als $a_1 \cdot q_1$ vorkommt, so ist δ größer (kleiner) als Null und es liegt positive (negative) Assoziation des Allels A_1 mit dem Allel Q_1 vor.

Methoden

Lange Zeit galten Fall-Kontroll-Studien für das Auffinden von allelischer Assoziation als das geeignete Mittel. Der Nachteil dieses Studientypes besteht in dem eventuell auftretenden Stratifikationseffekt, der bei einer ungeigneten Kontrollgruppe zutage tritt. Analysiert werde beispielsweise ein Marker, dessen Allele die Haarfarbe bestimmen. Ist nun der Anteil dunkelhaariger Menschen in der Fallgruppe höher als in der Kontrollgruppe, so wird man das Allel, welches für dunkle Haare kodiert, für krankheits(mit)verursachend halten. Um solche künstliche Assoziation zu vermeiden sollten sämliche Personen, die in der Untersuchung sind, aus einer möglichst homogenen Population stammen. Außerdem sollten Fall- und Kontrollgruppe für Variablen (Confounder) wie Alter, Geschlecht und ethnische Zugehörigkeit möglichst ähnlich strukturiert sein.

2.2 Transmission/Disequilibrium Test (TDT)

2.2.1 TDT für *einen* biallelischen Genort

Das im Kapitel 2.1 beschriebene Stratifikationsproblem besteht nicht bei familienbasierten Assoziationstests wie dem Transmission/Disequilibrium Test (TDT) von Spielman et al. (1993). In dieser ursprünglichen Version berücksichtigt der TDT nur Trios, also jeweils Vater und Mutter, deren Krankheitsstatus unerheblich ist, mit einem erkrankten Kind. Die Nullhypothese dieses Testes lautet

H_0: *Keine Kopplung oder keine Assoziation liegt vor.*

und somit ist die Alternativhypothese

H_1: *Sowohl Kopplung als auch Assoziation liegen vor.*

Im Falle eines biallelischen Markers liefert jedes Trio zwei Einträge in die folgende Tabelle:

Tabelle 2.1: Beobachtete Übertragungshäufigkeit bei *einem* biallelischen Marker

Übertragenes Allel	Nichtübertragenes Allel		Gesamt
	A_1	A_2	
A_1	t_{11}	t_{12}	$t_{11}+ t_{12}$
A_2	t_{21}	t_{22}	$t_{21}+ t_{22}$
Gesamt	$t_{11}+ t_{21}$	$t_{12}+ t_{22}$	$2n$

Die für den folgenden Test relevanten Tabelleneinträge b und c der Tabelle 2.1 zeigen, wieviele heterozygote Eltern das Allel A_1 und wieviele heterozygote Eltern das Allel A_2 auf das Kind vererbt haben. Nicht im Test berücksichtigt werden dagegen die Einträge t_{11} und t_{22}, da diese die Vererbung durch die

homozygoten Eltern beschreiben. Falls der Elter den homozygoten Genotyp (A_1,A_1) trägt, kann er nur Allel A_1 und, falls der Elter den anderen homozygoten Genotyp (A_2,A_2) trägt, kann der Elter nur das Allel A_2 übertragen. Die Information, dass ein homozygoter Elter das Allel A_1 oder A_2 übertragen hat ist für die untersuchte Fragestellung wertlos.

Beispiel

Zunächst wird an einem einfachen Beispiel die Durchführung des TDT demonstriert. Trios sollen an den Genorten A und B jeweils auf Assoziation untersucht werden. Gezeigt wird, welchen Beitrag das folgende Trio jeweils für die Genorte *A* und *B* liefert.

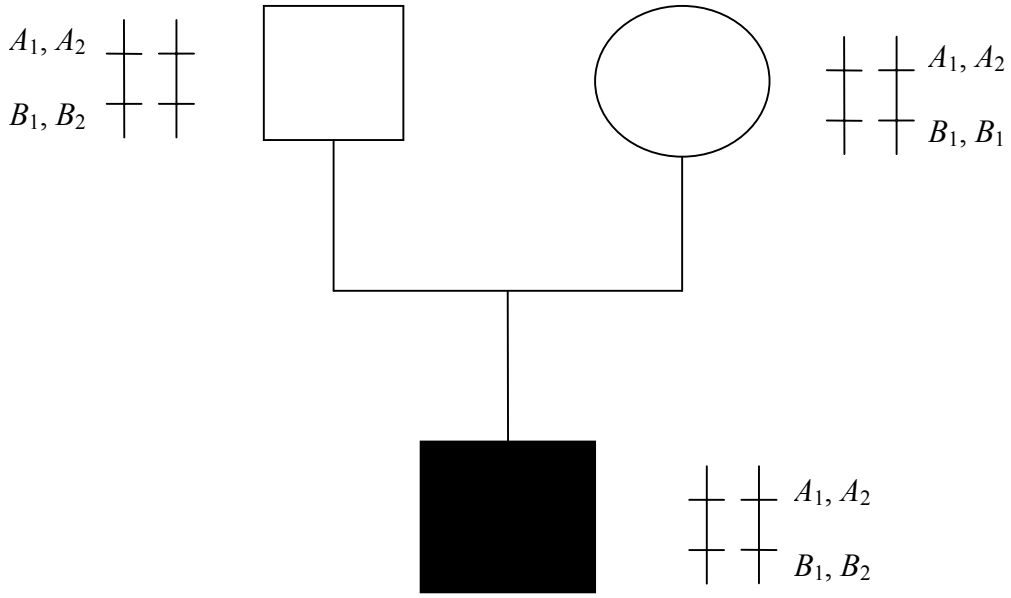


Abbildung 2.1: Trio mit zwei biallelischen Genorten

Am Genort *B* haben der Vater und das Kind den heterozygoten Genotyp (B_1,B_2) die Mutter hat den homozygoten Genotyp (B_1,B_1). Somit lässt sich eindeutig feststellen, dass der Vater das Allel B_2 übertragen und das Allel B_1 nicht übertragen hat und die Mutter ein Allel B_1 übertragen und das zweite Allel B_1

nicht übertragen hat. Diese Information ist in folgender Tabelle 2.2 zusammengefasst.

Tabelle 2.2: Übertragung von Allelen am Genort B im Trio der Abb. 2.1

nicht übertragenes Markerallel / über-tragenes Markerallel	B_1	B_2	Gesamt
B_1	t_{11}=1	t_{12}=0	1
B_2	t_{21}=1	t_{22}=0	1
Gesamt	2	0	2

Am Genort A ist nicht eindeutig, welcher Elter welches Allel vererbt hat: Entweder hat der Vater das Allel A_1 übertragen und Allel A_2 nicht übertragen und somit die Mutter das Allel A_2 übertragen und das Allel A_1 nicht übertragen. Oder aber der Vater hat das Allel A_2 übertragen und die Mutter hat das Allel A_1 übertragen. Trotz dieser Zweideutigkeit ist diese Situation für die Auswertung mit dem TDT unproblematisch, da beide möglichen Erklärungen zu denselben Tabelleneinträgen führen: Jeweils ist von einem Elter ein Allel A_1 übertragen worden und ein Allel A_2 nicht übertragen worden, von dem anderen Elter ist dagegen ein Allel A_2 übertragen worden und ein Allel A_1 nicht übertragen worden. Folglich erhöhen sich die Werte t_{12} und t_{21} jeweils um Eins. Solange man keine geschlechtsspezifischen Untersuchungen macht, ist es für die Kartierung von Krankheitsgenen unerheblich, welche dieser beiden Möglichkeiten vorliegt.

Testgröße und deren Verteilung

Die TDT-Teststatistik entspricht einer ursprünglich von McNemar (1947) entwickelten Testgröße zum Test auf nichtzufällige Abweichungen zwischen t_{12} und t_{21} in der Vierfeldertafel und lautet:

$$\hat{\chi}^2 = (t_{12}-t_{21})^2/(t_{12}+t_{21}) \qquad (2.1)$$

Diese Testgröße ist unter der Nullhypothese asymptotisch χ^2 -verteilt mit einem Freiheitsgrad. Da das 95%-Quantil dieser Verteilung 3.84 beträgt, erhält man einen Nachweis für Assoziation zu einem Signifikanzniveau von 5%, falls $\hat{\chi}^2 > 3.84$.

Vor- und Nachteile des Testverfahrens

Der TDT hat die Vorteile, dass er kein bestimmtes Vererbungsmodell voraussetzt, einfach durchführbar ist und, da interne Kontrollen verwendet werden und der Test somit familienbasiert ist, kein Stratifikationseffekt auftritt. Allerdings wird bei diesem Test der Genotyp der Eltern des erkrankten Kindes als bekannt vorausgesetzt, was bei Krankheiten mit einem hohen Erkrankungsalter nur selten gegeben sein dürfte.

Sind die elterlichen Genotypen nicht oder nur teilweise vorhanden, kann die Assoziationsuntersuchung eines Genortes mit einer Methode aus dem Kapitel 2.3 durchgeführt werden.

2.2.2 TDT für *einen* multiallelischen Genort

In diesem Kapitel wird gezeigt, wie man bei einem multiallelischen Genort auf Assoziation testet. Die TDT-Testgröße (2.1) kann verwendet werden, wenn a-priori ein Allel (z.B. Allel 1) als stark assoziiert vermutet wird, so dass die

übrigen Allele zu einer Gruppe "Allel 2" zusammengefasst werden können. Mit diesem dann "biallelischen" Genort testet man wie in Kapitel 2.2.1 beschrieben.

Direkte Erweiterung des TDT

Ohne a-priori-Annahmen lautet die von Spielman und Ewens (1996) vorgeschlagene direkte Erweiterung von (2.1) auf einen multiallelischen Genort:

$$TDT_k = \frac{(k-1)}{k} \cdot \text{Summe}_{i=1}^{k}(t_{i.} - t_{j.})^2/(t_{i.} + t_{j.} - 2t_{ii}) \qquad (2.2)$$

Der Index k bezeichnet die Anzahl der Allele an dem untersuchten Genort, $t_{i.}$ ($t_{j.}$) ist die Anzahl der Eltern, die das Allel A_i (A_j) auf das Kind übertragen, und t_{ii} ist die Anzahl der Eltern mit dem Genotyp (A_i,A_i) (siehe Tabelle 2.3). Die Testgröße (2.2) liefert einen Test der Nullhypothese

H_0: *Keine Kopplung oder keine allelische Assoziation liegt vor*.

Für k=2 erhält man die McNemar-Testgröße des TDT für biallelische Marker (2.1).

Die Größe TDT_k folgt nach Spielman und Ewens (1996) approximativ einer χ^2-Verteilung mit k-1 Freiheitsgraden.

Tabelle 2.3: Beobachtete Übertragungshäufigkeit bei *einem* biallelischen Marker

nicht über-tragenes Allel / über-tragenes Allel	A_1	A_2	...	A_k	Gesamt
A_1	t_{11}	t_{12}	...	t_{1k}	$t_{1.}$
A_2	t_{21}	t_{22}	...	t_{2k}	$t_{2.}$
:	:	:	. . .	:	:
A_k	t_{k1}	t_{k2}	...	t_{kk}	$t_{k.}$
Gesamt	$t_{.1}$	$t_{.2}$	...	$t_{.k}$	$t_{..}$

Beispiele

Die Auswertung von Daten soll an folgenden drei Trios demonstriert werden:

$$\frac{A_1}{A_2} \times \frac{A_2}{A_1} \rightarrow \frac{A_1}{A_2}$$

$$\frac{A_3}{A_1} \times \frac{A_3}{A_2} \rightarrow \frac{A_3}{A_3}$$

$$\frac{A_2}{A_1} \times \frac{A_3}{A_1} \rightarrow \frac{A_2}{A_3}$$

Tabelle 2.4: Beobachtete Übertragungshäufigkeit bei *einem* Marker mit drei Allelen (Beispiel)

nicht übertragenes Allel / übertragenes Allel	A_1	A_2	A_3	Gesamt
A_1	t_{11}=0	t_{12}=1	t_{13}=0	$t_{1.}$=1
A_2	t_{21}=2	t_{22}=0	t_{23}=0	$t_{2.}$=2
A_3	t_{31}=2	t_{32}=1	t_{33}=0	$t_{3.}$=3
Gesamt	$t_{.1}$=4	$t_{.2}$=2	$t_{.3}$=0	$t_{..}$=6

Den Wert der Testgröße erhält man dann durch:

$$TDT_3 = \frac{(3-1)}{3} \cdot \text{Summe}_{i=1}^{3}(t_{i.} - t_{.i})^2/(t_{i.} + t_{.i} - 2t_{ii})$$

$$= \frac{2}{3}\left| \frac{(1-4)^2}{(1+4-0)} + \frac{(2-2)^2}{(2+2-0)} + \frac{(3-0)^2}{(3+0-0)} \right|$$

$$= \frac{2}{3}\left| \frac{9}{5} + \frac{0}{4} + \frac{9}{3} \right|$$

$$= 3.2$$

Der Vergleich mit dem entsprechenden kritischen Wert, dem 95%-Quantil der χ^2-Verteilung ist aufgrund der Asymptotik nur bei einer größeren Anzahl von Trios sinnvoll.

Folgendes Beispiel zeigt einen Fall, für den die Testgröße auch bei einer großen Datenanzahl von der χ^2-Verteilung mit k-1 Freiheitsgraden abweicht. Man gehe von zwei Populationen aus, von denen die erste am Genort A nur die beiden Allele A_1 und A_2 aufweise und die zweite Population die beiden Allele A_3 und A_4. Diese beiden Populationen seien geografisch getrennt, so dass sich die Personen nur innerhalb ihrer Population fortpflanzen. Somit kommen keine Personen mit dem Genotyp (A_2,A_3) vor und es gilt:

$$t_{23}=t_{32}=0.$$

Die Testgröße TDT_k lässt sich dann folgendermaßen berechnen (k=4):

$$TDT_4 = \frac{3}{2}\left| \frac{(t_{12}-t_{21})^2}{t_{12}+t_{21}} + \frac{(t_{34}-t_{43})^2}{t_{34}+t_{43}} \right|.$$

Die beiden Summanden in der eckigen Klammer sind jeweils approximativ χ^2-verteilt mit einem Freiheitsgrad. Aus der Unabhängigkeit der beiden Summanden (denn es liegen zwei unabhängige Populationen vor) folgt, dass die Summe χ^2-verteilt ist mit zwei Freiheitgraden. Der Erwartungswert beträgt also zwei und die Varianz beträgt vier (Müller PH, 1991). Berücksichtigt man noch den Faktor $\frac{3}{4}$, wird der Erwartungswert $\frac{3}{2}$ und die Varianz beträgt $\frac{9}{4}$.

Für den Erwartungswert der Teststatistik erhält man drei und für die Varianz erhält man sechs $(2\cdot(k-1)=6)$, wenn man sie nach Spielman und Ewens (1996) bestimmt. Simulationen mit dem Software-Paket SAS (SAS Institute Inc., 1990) zeigen, dass durch diese in der vorliegenden Situation fehlerhaften Werte die Größe des Fehlers erster Art von 5% auf 0.55% sinkt. Somit hält der

Test das vorgegebene Niveau, da aber eine Absenken des Fehlers erster Art in der Regel zu einer Erhöhung des Fehlers zweiter Art führt, hat der Test eine verringerte Power.

Im Folgenden wird ein alternatives Modell im Rahmen einer Logistischen Regression erstellt. Das bedeutet, dass der Logarithmus der Odds (Wahrscheinlichkeit geteilt durch die Gegenwahrscheinlichkeit) modelliert wird (siehe Gleichung (2.3)). Gerechtfertigt ist dieser Ansatz durch Modellierung der theoretischen Übertragungswahrscheinlichkeiten im Falle *eines* Krankheitsgenortes. Für diese Modellierung setzt man die Unabhängigkeit der elterlichen Beiträge und zufällige Partnerwahl (random mating) voraus.

Erweiterung des TDT (ETDT) nach Sham und Curtis (1995)

Man betrachte die Vererbung der elterlichen Allele am Genort A auf ein krankes Kind. Hat der Vater den Genotyp (A_r,A_s) und die Mutter den Genotyp (A_t,A_u), dann sei $p_{rs,tu}$ die bedingte Wahrscheinlichkeit, dass ein krankes Kind vom Vater das Allel A_r und von der Mutter das Allel A_t vererbt bekommt (und die Allele A_s und A_u nicht). Knapp (1993) zeigt, sind die Beiträge der Eltern unter der Alternativhypothese im Allgemeinen nicht unabhängig voneinander. Jedoch gilt die Unabhängigkeit unter der Nullhypothese

H_0*: Keine Kopplung oder keine Assoziation liegt vor.*

Somit lässt sich die Übertragungswahrscheinlichkeit für beide Eltern in zwei Faktoren aufspalten

$$p_{rs,tu} = p_{rs} \cdot p_{tu},$$

wobei p_{ij} die Wahrscheinlichkeit ist, dass ein Elter das Allel A_i auf ein bestimmtes Kind überträgt und Allel A_j somit nicht überträgt unter der Bedingung, dass der elterliche Genotyp (A_i,A_j) lautet und das Kind krank ist. Daraus folgt direkt: $p_{ij}=1- p_{ji}$.

Notation

Außerdem definiert man:

A	Marker mit k Allelen
A_i	Allel am Marker A
Q	Krankheitsgenort mit zwei Allelen
Q_1	Krankheitsallel am Krankheitsgenort Q
Q_2	normales Allel am Krankheitsgenort Q
p_i	Häufigkeit des Allels A_i am Marker A
q_s	Häufigkeit des Allels Q_s am Krankheitsgenort Q
h_{si}	Häufigkeit des Haplotyps Q_sA_i
f_{st}	Penetranz des Genotyps (Q_s,Q_t) am Krankheitsgenort Q (Erkrankungswahrscheinlichkeit einer Person mit dem Genotyp (Q_s,Q_t)

und

$$e_{si} := \frac{h_{si}}{q_s p_i}$$

$$B := \left\{ q_1 \left[q_1 (f_{11} - f_{12}) + q_2 (f_{12} - f_{22}) \right] \right\} \Big/ \left\{ q_1^2 f_{11} + 2 q_1 q_2 f_{12} + q_2^2 f_{22} \right\}$$

sowie $\quad d_{ij} := 1 + B \cdot \left[\left(e_{1i} - 1 \right) + \theta \left(e_{1j} - e_{1i} \right) \right].$

Die drei zuletzt definierten Größen lassen sich wie folgt interpretieren: e_{si} beschreibt die Stärke der Assoziation zwischen den Allelen A_i und Q_s, B ist ein Maß dafür, wie sehr das normale Allel Q_2 am Krankheitsgenort durch Selektion aus den kranken Kindern verdrängt worden ist, und d_{ij} steht für das Ausmaß ungleicher Übertragung der Allele A_i und A_j durch einen Elter mit dem Genotyp (A_i,A_j).

Tabelle 2.5: Übertragungswahrscheinlichkeiten bei *einem* Krankheitsgenort in einer Untersuchung von Trios

nicht übertragenes Allel / übertragenes Allel	A_1	A_2	$\cdots$	A_k	Gesamt
A_1	$p_1^2 d_{11}$	$p_1 p_2 d_{12}$	$\cdots$	$p_1 p_k d_{1k}$	$p_1\left[1+B\cdot(1-\theta)\cdot(e_{11}-1)\right]$
A_2	$p_2 p_1 d_{21}$	$p_2^2 d_{22}$	$\cdots$	$p_2 p_k d_{2k}$	$p_1\left[1+B\cdot(1-\theta)\cdot(e_{11}-1)\right]$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	$p_k p_1 d_{k1}$	$p_k p_2 d_{k2}$	$\cdots$	$p_k^2 d_{kk}$	$p_1\left[1+B\cdot(1-\theta)\cdot(e_{11}-1)\right]$
Gesamt	$p_1[1+B\theta$ $(e_{11}-1)]$	$p_2[1+B\theta$ $(e_{12}-1)]$	$\cdots$	$p_k[1+B\theta$ $(e_{1k}-1)]$	1

Herleitung des Regressionsmodells

Aus der Definition der Größe d_{ij} folgt, dass d_{ij} nur für $\theta>0$ von j abhängt. Dies ermöglicht Sham und Curtis (1995) zu zeigen, dass gilt:

$$p_{ij} = \frac{P(\text{Kind ist krank})\cdot p_i p_j d_{ij}}{P(\text{Kind ist krank})\cdot p_i p_j \left(d_{ij}+d_{ji}\right)} = \frac{d_{ij}}{d_{ij}+d_{ji}} \xrightarrow{\theta\to 0} \frac{d_{ii}}{d_{ii}+d_{jj}}.$$

Auf diese Weise erhält man folgende Modellgleichung der Logistischen Regression:

$$\ln\left|\frac{p_{ij}}{p_{ji}}\right| = \ln\left(\frac{\frac{d_{ii}}{d_{ii}+d_{jj}}}{\frac{d_{jj}}{d_{jj}+d_{ii}}}\right) = \ln\left(\frac{d_{ii}}{d_{jj}}\right) = \ln(d_{ii}) - \ln(d_{jj}) =: b_i - b_j \qquad (2.3)$$

Dieses ist äquivalent zu:

$$p_{ij} = \frac{\exp(b_i - b_j)}{1 + \exp(b_i - b_j)}$$

Im Modell sind also k Parameter b_1 bis b_k. Unter der Nullhypothese

$$H_0: \; b_1 = b_2 = ... = b_k$$

sind die p_{ij} gleich 0.5 für alle i,j. Auf der linken Seite der Gleichung (2.3) steht dann eine Null. Die Gleichung ist also genau dann erfüllt, wenn alle Parameter b_1 bis b_k identisch sind.

Durchführung des Testes

Man möchte wissen, wie stark die Daten von der Nullhypothese abweichen. Dazu schätzt man die Parameter durch Maximierung der Log-Likelihood

$$L = \text{Summe}_{i<j} \left\{ t_{ij} \ln\left(p_{ij}\right) + t_{ji} \ln(p_{ji}) \right\}$$

unter der Alternativhypothese. Um die Parameter eindeutig schätzen zu können wird b_k auf Null gesetzt. Diese maximierte Log-Likelihood sei mit L_1 bezeichnet Zusätzlich zu L_1 benötigt man noch die Größe der Likelihood unter der Nullhypothese (L_0).

Der Vergleich von L_0 mit L_1 erfolgt durch die Likelihood-Ratio-Testgröße

$$G_1 = 2 \cdot (L_1 - L_0).$$

Diese ist χ^2-verteilt mit k-1 Freiheitsgraden. Ist also die Teststatistik größer als das (1-α)-Quantil der entsprechen Verteilung, wird die Nullhypothese abgelehnt. Andernfalls kann die Nullhypothese nicht verworfen werden, d.h. zum Niveau von α kann keine Assoziation nachgewiesen werden. Somit steht

ein Testverfahren mit wenigen Parametern und somit in der Regel einer großen Power zur Verfügung.

Beispiel

Gegeben seien wieder die drei Trios:

$$\frac{A_1}{A_2} \times \frac{A_2}{A_1} \rightarrow \frac{A_1}{A_2}$$

$$\frac{A_3}{A_1} \times \frac{A_3}{A_2} \rightarrow \frac{A_3}{A_3}$$

$$\frac{A_2}{A_1} \times \frac{A_3}{A_1} \rightarrow \frac{A_2}{A_3}$$

Die Likelihood dieser Daten lautet:

$$L = \text{Summe}_{i<j} \left\{ t_{ij} \ln\left(p_{ij}\right) + t_{ji} \ln(p_{ji}) \right\}$$

$$= t_{12} \cdot \ln(p_{12}) + t_{21} \cdot \ln(p_{21})$$

$$+ t_{13} \cdot \ln(p_{13}) + t_{31} \cdot \ln(p_{31})$$

$$+ t_{23} \cdot \ln(p_{23}) + t_{32} \cdot \ln(p_{32})$$

$$= 1 \cdot \ln(p_{12}) + 2 \cdot \ln(p_{21})$$

$$+ 0 \cdot \ln(p_{13}) + 2 \cdot \ln(p_{31})$$

$$+ 0 \cdot \ln(p_{23}) + 1 \cdot \ln(p_{32})$$

Unter der Nullhypothese gilt p_{ij}=0.5 für i,j=1,2,3. Daraus folgt:

$$L_0 = 6 \cdot \ln(0.5) = -4.16$$

Die maximale Likelihood erhält man für b_1=-135.65, b_2=-134.96 und b_3=0. Es gilt:

$$L_1 = 1 \cdot \ln \frac{\exp(-135.65 - 134.96)}{1 + \exp(-135.65 - 134.96)} + 2 \cdot \ln \frac{\exp(-134.96 - 135.65)}{1 + \exp(-134.96 - 135.65)}$$

$$+ 0 \cdot \ln \frac{\exp(-135.65 - 0)}{1 + \exp(-135.65 - 0)} + 2 \cdot \ln \frac{\exp(0 - 135.65)}{1 + \exp(0 - 135.65)}$$

$$+ 0 \cdot \ln \frac{\exp(-134.96 - 0)}{1 + \exp(-134.96 - 0)} + 1 \cdot \ln \frac{\exp(0 - 134.96)}{1 + \exp(0 - 134.96)}$$

$$= -1.91$$

und somit

$$G_1 = 2 \cdot (L_1 - L_0) = -1.91 - (-4.16) = 4.5.$$

Der Vergleich mit den Quantilen der χ^2-Verteilung mit zwei Freiheitsgraden ergibt einen p-Wert von 0.11. Die Nullhypothese wird zum Signifikanzniveau von 0.05 nicht verworfen, Assoziation kann also nicht gefolgert werden.

2.3 Familienbasierte Assoziationsanalyse bei fehlenden Eltern

Insbesondere dann, wenn man erst spät auftretende Krankheiten wie die Alzheimer-Krankheit oder Krebs untersucht, wird es nur in Ausnahmefällen möglich sein, die Genotypen der Eltern des Erkrankten zu erhalten. Fehlen diese, kann nicht ohne Weiteres festgestellt werden, welche Allele übertragen

und welche Allele nicht übertragen worden sind. Da dieses Wissen für die Durchführung des TDT unentbehrlich ist, muss für einen Test auf Assoziation ein alternatives Verfahren verwendet werden. Ein solches Verfahren nutzt die Information in den Genotypen der Geschwister des Erkrankten entweder direkt, indem man die Allelhäufgkeiten in den kranken und gesunden Kindern miteinander vergleicht und bei hinreichend großer Abweichung auf Assoziation des Markers mit der Krankheit entscheidet (Kapitel 2.3.1 und 2.3.2), oder man rekonstruiert die elterlichen Genotypen mit Hilfe der Genotypen der Kinder (Kapitel 2.3.3).
Auf diese Weise sind auch bei fehlenden elterlichen Genotypen TDT-ähnliche Untersuchungen möglich. Drei Verfahren sollen hier vorgestellt werden.

2.3.1 Sib Transmission/Disequilibrium Test (S-TDT)

Der Sib Transmission/Disequilibrium Test (S-TDT) von Spielman und Ewens (1998) ist ein Test der Nullhypothese

H_0*: Marker und Krankheitsgenort sind nicht gekoppelt* (θ*=0.5).*

Für den S-TDT können, anders als beim TDT, auch solche Familien berücksichtigt werden, bei denen keine elterlichen Genotypen vorliegen. Allerdings wird dann mindestens ein gesundes und ein krankes Kind pro Geschwisterschaft vorausgesetzt. Außerdem dürfen nicht alle Kinder denselben Genotyp aufweisen. Die "ideale" Nullhypothese des TDT

H_0: *Keine Kopplung oder keine Assoziation liegt vor*

kann mit dem S-TDT nur getestet werden, wenn sämtliche Familien in der Studie die Minimalkonfiguration aufweisen, also genau ein krankes und genau ein gesundes Kind in jeder Familie vorkommt. Für die Durchführung des Testes gibt es zwei Vorgehensweisen, die im Folgenden beschrieben werden. Liegt eine kleine Anzahl an Familien vor, ist die Permutations-Methode die angemessene Vorgehensweise, bei einer großen Anzahl von Familien

verwendet man die z-Score-Methode. Letztere lässt dann auch eine Kombination der Information aus den Geschwisterschaften mit der Information aus eventuell vorliegenden Trios zu.

Durchführung des Testes nach der Permutationsmethode

Allel A_1 sei das Allel, für welches man auf Assoziation testen möchte. Die Permutations-Methode weist jedem der k kranken und g gesunden Kinder je Geschwisterschaft zufällig einen Gesundheitszustand zu, so dass die Anzahl der kranken und gesunden Kinder in jeder Familie unverändert bleibt. Diese Zuweisung wiederholt man und bestimmt die Summe der Anzahlen des A_1-Allels in den Geschwisterschaften. Der p-Wert des Testes ergibt sich dann aus dem Anteil der Wiederholungen, bei denen man eine mindestens ebenso unwahrscheinliche Verteilung des A_1-Allels erhält wie in der tatsächlichen Population.

Beispiel

Gegeben seien folgende drei Geschwisterschaften mit jeweils einem gesunden und einem kranken Kind:

Tabelle 2.6: Genotyp-Verteilung in drei Geschwisterschaften mit jeweils einem gesunden und einem kranken Kind

Geschwister-schaft	Krankheits-zustand	(A_1,A_1)	(A_1,A_2)	(A_1,A_3)	(A_2,A_2)
1	krank	1	0	0	0
	gesund	0	1	0	0
2	krank	0	1	0	0
	gesund	0	0	1	0
3	krank	1	0	0	0
	gesund	0	0	0	1

Offenbar gibt es in den kranken Kindern insgesamt fünf Mal das Allel A_1. Simulationen ergeben, dass man durch Permutation des Krankheitszustandes innerhalb der Geschwisterschaften in 25% der Fälle ebenfalls fünf Allele A_1 in den kranken Kindern erhält. Da dieser Anteil höher als das Signifikanzniveau von 5% liegt, wird die Nullhypothese nicht abgelehnt.

Durchführung des Testes nach der z-Score-Methode

Bei der z-Score-Methode vergleicht man die beobachtete Anzahl Y des Allels A_1 in den erkrankten Geschwistern des Datensatzes mit der erwarteten Anzahl

$$E(Y) = \text{Summe} \frac{(2r+s)k}{t},$$

wobei r die Anzahl der Geschwister mit Genotyp (A_1,A_1) ist, s die Anzahl der Geschwister mit dem heterozygoten Genotyp (A_1,A_2) und t die Anzahl aller Geschwister, jeweils pro Familie, bezeichnet. Summiert wird über die Familien. Der N(0,1)-verteilte z-Score berücksichtigt außerdem die Varianz der Allelanzahl

$$Var(Y) = \text{Summe} \frac{k \cdot g\left[4r(t-r-s)+s(t-s)\right]}{t^2(t-1)}$$

und eine Stetigkeitskorrektur von 0.5.

Den p-Wert erhält man dann als Wahrscheinlichkeit, mit der eine N(0,1)-verteilte Zufallsvariable größer ist als

$$z' = \frac{\left(|Y - E(Y)| - \frac{1}{2}\right)}{\sqrt{Var(Y)}}.$$

Möchte man zusätzlich auch Familien auswerten, bei denen beide Eltern typisiert sind, bestimmt man die Anzahl (X) der übertragenen A_1-Allele von den heterozygoten Eltern auf kranke Kinder. Weil gilt:

$$E(X)=\frac{n}{2}$$

und

$$Var(X)=\frac{n}{4}.$$

lautet die kombinierte Testgröße

$$z_{comb}=\frac{\left|X+Y-\frac{n}{2}-Var(Y)\right|-\frac{1}{2}}{\sqrt{\frac{n}{4}+Var(Y)}}.$$

Der Vergleich mit den Quantilen der N(0,1)-Verteilung führt dann zu dem p-Wert für den Test der kombinierten Stichprobe.

2.3.2 Sibship Disequilibrium Test (SDT)

Der von Horvath und Laird (1998) entwickelte Sibship Disequilibrium Test (SDT) testet im Gegensatz zum S-TDT auch für größere Familien die Nullhypothese

H_0: *Keine Kopplung oder keine Assoziation liegen vor*

gegen die Alternative

H_1: *Kopplung oder Assoziation liegen vor*.

Für den Test benötigt man Geschwisterschaften mit mindestens einem kranken und mindestens einem gesunden Kind. Die Idee des SDT liegt darin, die Allelhäufigkeiten bei kranken und gesunden Kindern miteinander zu vergleichen und bei einer großen Differenz in vielen Geschwisterschaften (mittels des Vorzeichentestes) allelische Assoziation zu folgern. Zunächst wird die Vorgehensweise bei Vorliegen *eines* biallelischen Markers beschrieben.

SDT bei *einem* biallelischen Marker

Bezeichne b die Anzahl der Familien, in denen ein bestimmtes Allel (etwa Allel A_1) in den kranken Kindern häufiger auftritt als in den gesunden Kindern und c sei die Anzahl der Familien, in denen dieses Allel bei den gesunden Kindern häufiger auftritt als bei den kranken. Unter H_0 folgt die Testgröße b einer Binomialverteilung mit den Parametern $b+c$ und 0.5.

Beispiel

In der Situation von Tabelle 2.6 soll die Bestimmung von b und c demonstriert werden (Allele A_2 und A_3 werden zu einem Allel zusammengefaßt). Es gibt zwei Geschwisterschaften in denen bei den erkrankten Kinder mehr Allele A_1 vorkommen als in den gesunden Kinder: In Geschwisterschaft eins gibt es zwei Allele A_1 im erkrankten Kind und nur ein Allel A_1 im gesunden Kind. In Geschwisterschaft zwei taucht das Alllel A_1 jeweils ein Mal im kranken und gesunden Kind auf und in Geschwisterschaft drei gibt es zwei Allele A_1 im kranken Kind und kein Allel A_1 im gesunden Kind. Also gilt für dieses Beispiel:

$$b=2 \text{ und } c=0.$$

Testgröße im Fall vieler Familien

Liegen viele Familien vor, kann als Testgröße auch der Ausdruck

$$T = \frac{(b-c)^2}{b+c} \tag{2.6}$$

verwendet werden. Diese ist ein Maß für die Differenz zwischen der Anzahl der Familien, in denen das untersuchte Allel häufig auftritt und der Anzahl der Familien, bei denen das untersuchte Allel selten ist. Bei Vorliegen eines großen Stichprobenumfanges folgt diese Testgröße einer χ^2-Verteilung mit einem Freiheitgrad.

SDT bei *einem* multiallelischen Marker

Der SDT für *einen* multiallelischen Marker entspricht einem multivariaten Vorzeichentest der Differenzen d_j der relativen Allelhäufigkeiten in den gesunden und in den kranken Kindern, mit j=1,...,m, wobei m die Anzahl der Allele am Marker bezeichnet. Außerdem sei S_j die Anzahl der Geschwisterschaften, in denen die relative Häufigkeit des Allels A_j in den erkrankten Kindern höher ist als in den gesunden, vermindert um die Anzahl der Geschwisterschaften, in denen die relative Häufigkeit des Allels A_j in den gesunden Kindern höher ist als in den kranken Kindern. Diese Werte werden im Vektor S zusammengefasst:

$S'=(S_1, S_2, ..., S_{m-1})$. Schließlich benötigt man noch die Matrix W, deren Elemente W_{jk} wie folgt definiert sind. Für jedes Allelpaar A_j und A_k wird jeder Geschwisterschaft der Wert 1 zugewiesen, wenn für diese gilt, dass d_j und d_k das gleiche Vorzeichen haben, sonst -1. Aufsummiert über die Familien ergibt sich der Wert W_{jk}. Die Testgröße des multivariaten Vorzeichen-Testes $T = S'W^{-1}S$ ist asymptotisch χ^2-verteilt mit m-1 Freiheitsgraden. Berechnet

man diese Testgröße bei einem biallelischen Marker, erhält man die oben beschriebene Testgröße (2.6).

2.3.3 Reconstruction combined TDT (RC-TDT)

Eine weitere Möglichkeit für die Vorgehensweise bei ganz oder teilweise fehlenden elterlichen Genotypinformationen beschreibt Knapp (1999). Anders als bei den oben beschriebenen Verfahren S-TDT und SDT, werden beim RC-TDT die elterlichen Genotypen, soweit dies möglich ist, anhand der Genotypen der Kinder rekonstruiert. Für die Durchführung des Testes ist mit Rekonstruktion nicht die exakte Bestimmung der Genotypkonstellation gemeint. Man muss lediglich entscheiden können, ob die Eltern für das interessierende Allel homozygot oder heterozygot sind oder ob das Allel nicht im Genotyp vorkommt. Nach der Rekonstruktion kann die Familie nicht ohne Weiteres für den TDT verwendet werden, da dies zu einer Erhöhung des Fehlers erster Art führen kann. Folgende Vorgehensweise erlaubt einen asymptotischen Test zum Signifikanzniveau α.

Durchführung des Testes

In Familien, bei denen mindestens ein elterlicher Genotyp nicht rekonstruiert werden kann, werden Erwartungswert und Varianz der Anzahl des interessierenden Allels, welches mit A_1 bezeichnet sei, in erkrankten Kindern bestimmt.

Urnenmodell

Die Anzahlen der erkrankten Kinder, die homozygot für das Allel A_1 sind (x), und solcher erkrankter Kinder, die heterozygot für dieses Allel sind (y), kann man sich als Anzahl gezogener Kugeln mit einer "1" bzw "0" aus einer Urne vorstellen. Diese Urne enthalte t (=Gesamtzahl der Kinder, die homozygot für

A_1 oder heterozygot sind) Kugeln von denen r (=Anzahl der Kinder, die homozygot für A_1 sind) Kugeln durch eine "1" und die restlichen durch eine "2" markiert sind. Beim Ziehen ohne Zurücklegen folgen die Anzahl der mit "1" oder "2" markierten Kugeln jeweils einer hypergeometrischen Verteilung (siehe Tabelle 2.7). Somit kann ihr Erwartungswert und ihre Varianz durch Einsetzen in die entsprechenden Formeln bestimmt werden. Der Erwartungswert der Allelanzahl ist dann $2 \cdot E(x) + E(y)$ und die Varianz erhält man durch 4Var(x)+Var(y)+4Cov(x,y)(Spielman und Ewens, 1998).

Tabelle 2.7: Genotyphäufigkeiten in den Kindern

Genotypen	krank	gesund	Gesamt
(A_1,A_1)	x	r-x	r
(A_1,A_2)	y	s-y	s
Gesamt	a	u	t

In allen anderen Familien mit mindestens einem für das Allel A_1 heterozygoten Elternteil lassen sich Erwartungswert und Varianz der oben zitierten Veröffentlichung entnehmen. Für jede denkbare Kombination elterlicher Genotypen sind Erwartungswert und Varianz der Allelanzahl des Allels A_1 bestimmt worden. Wegen des großen Umfanges wird hier auf eine separate Aufführung aller Werte verzichtet. Sie sind der Veröffentlichung von Knapp (1999) zu entnehmen

Testgröße und derer Verteilung

Die Testgröße des RC-TDT lautet:

$$\frac{\text{Summe}_i\left(T_i - E(T_i)\right)}{\sqrt{\text{Summe}_i Var(T_i)}},$$

mit T_i:=Anzahl der Allele A_1 in den kranken Kindern. Unter der Nullhypothese

H_0: *Keine Kopplung oder keine Assoziation liegt vor*,

ist diese asymptotisch N(0,1)-verteilt.

Vergleich des RC-TDT mit dem S-TDT

Der RC-TDT wurde mittels Simulation mit dem S-TDT hinsichtlich der Power in Familien mit zwei, vier und sechs Geschwistern, von denen jeweils mindestens eines erkrankt ist verglichen. Dabei wurde jeweils drei Modelle (Penetranz bei zwei Krankheitsallelen f_{DD}=0.2, f_{DD}=0.5 und f_{DD}=0.8) mit dominantem, additiven und rezessiven Krankheitsmodell untersucht. Der RC-TDT ist dem S-TDT in allen untersuchten Modellen überlegen oder zumindest gleichwertig. Die Differenz in der Power liegt in den meisten Situationen unter 5% und liegt bis auf eine Ausnahme unter 10%. Die deutlichsten Unterschiede finden sich in allen untersuchten Familiengrößen bei dem additiven Modell. Eine gleichmäßige Abhängkeit von der Penetranz ist nicht erkennbar

Kapitel 3

Familienbasierte Assoziationsanalyse bei mehr als einem Marker

3.1 Problemstellung und Übersicht

In diesem Kapitel wird neben bereits bestehender Software ein neuer familienbasierter Test auf Assoziation für mehrere eng gekoppelte Marker vorgestellt. Ein solcher Assoziationstest, der mehrere Marker gleichzeitig berücksichtigt, ist insbesondere in Situationen wertvoll, in denen biallelische Marker vorliegen. Dann nämlich ist der einzelne Marker häufig aufgrund von Homozygotie uninformativ.

Homozygotie bei SNPs

Man betrachte beispielsweise einen biallelischen Marker A, den man als "single nucleotide polymorphism" (SNP) bezeichnet, mit den Allelen A_1 (mit der Häufigkeit p) und A_2 (mit der Häufigkeit q=1-p). Unter Annahme des Hardy-Weinberg-Gleichgewichts (Cavalli-Sforza und Bodmer, 1991) beträgt der Anteil heterozygoter Personen dann P(heterozygot)=2pq. Dieser Anteil wird maximal (nämlich 0.5) für p=q=0.5 (Abbildung 3.1). Das bedeutet, dass selbst im günstigsten Fall nur 50% aller Eltern informativ bezüglich der Übertragung an diesem Genort, also heterozygot, sind.

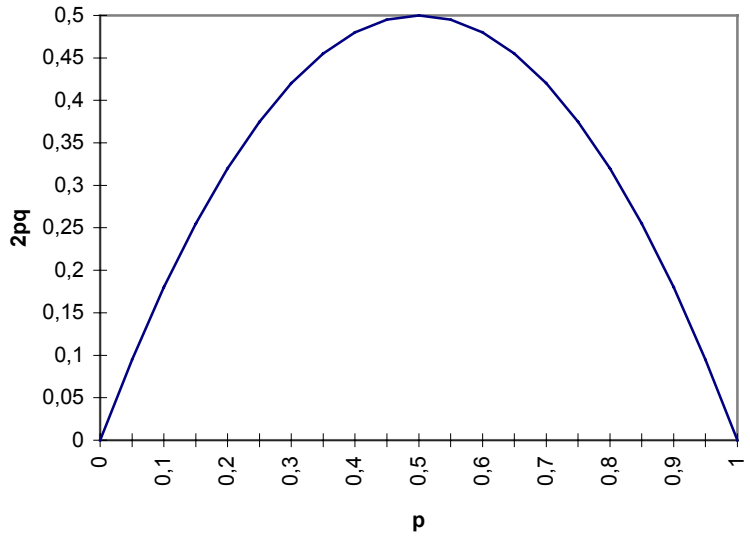


Abbildung 3.1: Heterozygotie bei SNPs

Zwei eng gekoppelte SNPs mit bekannter Phase kann man als einen Genort ("Superlocus") mit vier verschiedenen Allelen interpretieren. Bei beiden Sichtweisen gibt es 16 verschiedene Genotypen, wenn man berücksichtigt, ob ein Allel oder Haplotyp vom Vater oder von der Mutter geerbt wurde. Vier von diesen sind jeweils homozygot und damit nicht informativ bezüglich Assoziationsuntersuchungen.

Bei den zwei gekoppelten SNPs lauten die uninformativen Genotypen:

$$\frac{A_1B_1}{A_1B_1}, \frac{A_1B_2}{A_1B_2}, \frac{A_2B_1}{A_2B_1} \text{ und } \frac{A_2B_2}{A_2B_2},$$

bei dem einzelnen Genort mit vier Allelen:

$$\frac{A_1}{A_1}, \frac{A_2}{A_2}, \frac{A_3}{A_3} \text{ und } \frac{A_4}{A_4}.$$

Die Hinzunahme jedes weiteren biallelischen Markers halbiert den Anteil der an diesem "Superlocus" uninformativen Personen, denn die Anzahl der homozygoten Genotypen verdoppelt sich jeweils während die Gesamtanzahl der Genotypen sich jeweils vervierfacht. Falls die Haplotypen bekannt sind und die Rekombination vernachlässigt werden kann, entsprechen zwei SNPs hinsichtlich ihrer Information einem einzelnen Marker mit vier Allelen, drei SNPs entsprechen einem Marker mit acht Allelen und allgemein entsprechen t SNPs einem Marker mit 2^t Allelen.

Eindeutigkeit der Phase

Die Verwendung vieler SNPs erhöht allerdings die Wahrscheinlichkeit dafür, dass die Phase nicht eindeutig zu bestimmen ist und somit nicht eindeutig ist, welche Haplotypen von den Eltern auf das Kind vererbt wurden. Ein Trio ist in diesem Sinne "mehrdeutig", wenn die Eltern und das Kind an mindestens einem Genort gleichartig heterozygot sind und das Kind an mindestens zwei Genorten heterozygot ist (Abbildung 3.2).

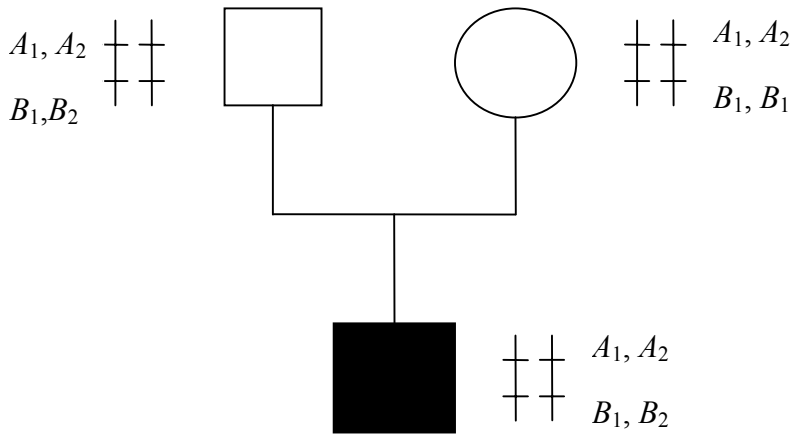


Abbildung 3.2: Nicht eindeutige Phase in einem Trio

Bezeichnet man die Allelhäufigkeiten an den k Genorten mit $p_1,...,p_k$ und $q_1,...,q_k$, dann erhält man für die Populationshäufigkeit bei Trios mit für das Kind eindeutigen Haplotypen (Hodge, 1999):

$$\begin{aligned} &\text{P(Haplotyp eindeutig)} \\ &= \prod_{i=1}^{k}\left(1-2p_i^2q_i^2\right) - \sum_{j=1}^{k}\left(2p_j^2q_j^2\right)\prod_{i\neq j}\left(1-2p_iq_i\right) \\ &\overset{p_i=const\ \forall\, i}{=} \left(1-2p^2q^2\right)^k - k\left(2p^2q^2\right)\left(1-2pq\right)^{k-1} \end{aligned}$$

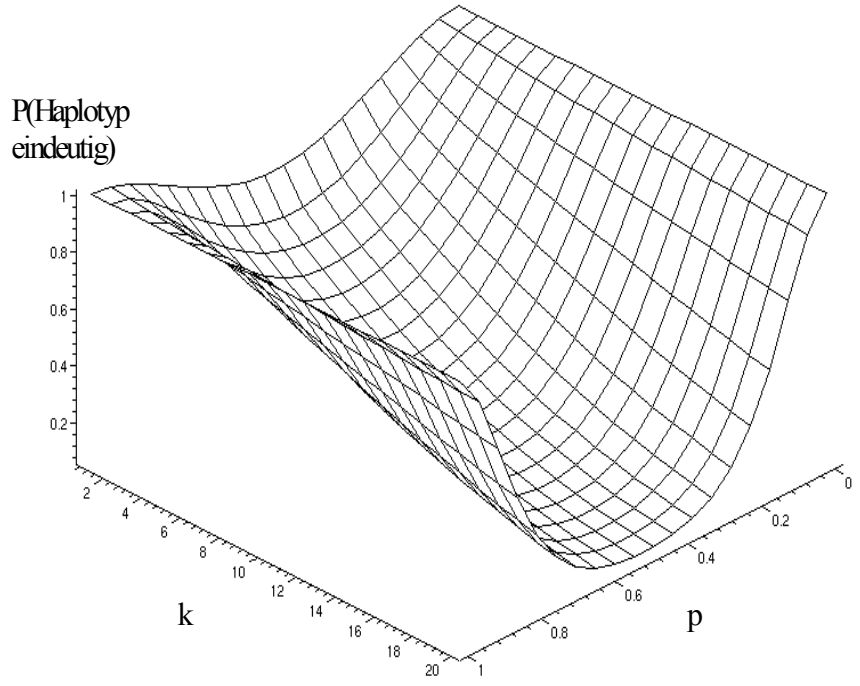


Abbildung 3.3: Populationshäufigkeit für Trios mit für das Kind eindeutigen Haplotypen

Offenbar ist die Wahrscheinlichkeit für eindeutige Haplotypen umso kleiner, je mehr Genorte betrachtet werden. Diese Beoachtung legt nahe, eine Methode zum Testen auf Assoziation zu verwenden, die im Gegensatz zur Methode von Wilson (1997) die Eindeutigkeit nicht voraussetzt (siehe Kapitel 3.3).

3.2 Bestehende Software

3.2.1 GENEHUNTER

GENEHUNTER (Kruglyak et al., 1996) ist ein Programm-Paket, welches parametrische und nichtparametrische Kopplungsanalyse mit vielen Markern durchführen kann. Dies ist besonders dann notwendig, wenn man die (mit)verursachenden Gene einer komplexen Krankheit grob kartieren möchte und Stammbaumdaten vorliegen.

Test auf Assoziation mit GENEHUNTER bei *einem* Marker

In der GENEHUNTER-Version 2.0 beta werden die Unterprogramme TDT, TDT2, TDT3 und TDT4 angeboten. Im Falle eines biallelischen Markers entspricht der von GENEHUNTER durchgeführte TDT dem von Spielman et al. (1993) beschriebenen TDT. Ist der Marker multiallelisch, wird jedes Allel gegen die restlichen in einer Gruppe zusammengefassten Allele getestet. Der Vergleich der McNemar-Testgröße (1947)

$$TDT = \frac{(b-c)^2}{b+c}$$

mit den Quantilen der χ^2-Verteilung führt dann zu den von GENEHUNTER angegebenen p-Werten. Da bei dieser Vorgehensweise viele Tests an einem

Datensatz durchgeführt werden, muss für den Nachweis einer Signifikanz zu einem Niveau von beispielsweise 5% ein p-Wert deutlich unterhalb der 5%-Marke angegeben sein. Ohne diese Korrektur werden zu viele Testergebnisse zufälligerweise signifikant.

Test auf Assoziation mit GENEHUNTER bei mehr als *einem* Marker

Die Kommandos TDT2, TDT3 und TDT4 berechnen Zwei-, Drei- und Vier-Genort-Versionen des TDT. Bei diesen wird gezählt, wie oft jeder mögliche Haplotyp von den Eltern auf das kranke Kind vererbt (diese Anzahl sei mit *b* bezeichnet) und wie oft er nicht vererbt wurde (diese Anzahl sei mit *c* bezeichnet).

Die Testgröße $TDT = \frac{(b-c)^2}{b+c}$ und der entsprechende p-Wert werden (unter der Annahme einer χ^2-Verteilung) berechnet. Sämtliche Familien, in denen an mindestens einem Genort beide Eltern gleichartig heterozygot sind und das Kind ebenfalls heterozygot ist, werden von GENEHUNTER nicht bei der Analyse berücksichtigt. Zwar ist dann die Phase unbekannt und somit die Übertragung der Haplotypen auf die Kinder nicht eindeutig, trotzdem enthalten solche Familien Informationen, die für einen Test auf Kopplung oder allelische Assoziation relevant sind. Eine solche Software, die nicht sämtliche Information nutzt, wird in der Regel vorhandene Assoziation seltener nachweisen können als eine Methode, die auch von GENEHUNTER nicht ausgewerteten Familien verwendet.

Anteil der von GENEHUNTER nicht ausgewerteten Familien

Am Beispiel von Trios mit biallelischen Markern soll im Folgenden gezeigt werden, wie hoch der Anteil der von GENEHUNTER nicht ausgewerteten Familien ist, wenn sämtliche Allelhäufigkeiten 0.5 betragen und keine Assoziation zwischen den Markerallelen vorliegt.

Es bezeichne *NB* das Ereignis, dass GENEHUNTER ein Trio aufgrund von gleichartiger Heterozygotie an mindestens einem Marker bei allen drei Triomitgliedern bei der Auswertung nicht berücksichtigt. Zunächst ist die Wahrscheinlichkeit 0.125, dass an dem interessierenden Marker beide Eltern und das Kind heterozygot sind. Daher gilt:

P(*NB*|zwei Genorte)=1-P(an keinem der beiden Marker sind beide Eltern und das Kind heterozygot)= $1-|1-\frac{1}{8}|^2=\frac{64-49}{64}=\frac{15}{64}\approx 0.23$.

Offenbar gilt für den Anteil nicht berücksichtigter Familien bei zwei Allelen und *k* Genorten

$$P(NB|k \text{ Genorte}) = 1-|\frac{7}{8}|^k.$$

Bei drei Allelen an jedem der *k* Genorte erhält man

$$P(NB|k \text{ Genorte}) = 1-|\frac{26}{27}|^k.$$

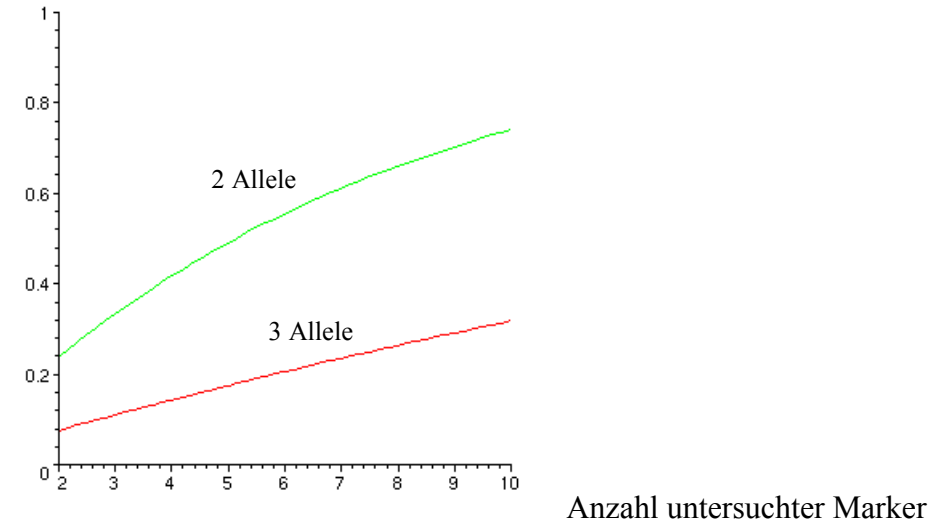


Abbildung 3.4: Anteil der Familien, die von GENEHUNTER bei der Auswertung durch TDT2, TDT3, TDT4 und entsprechende Programme für mehr Genorte nicht berücksichtigt werden

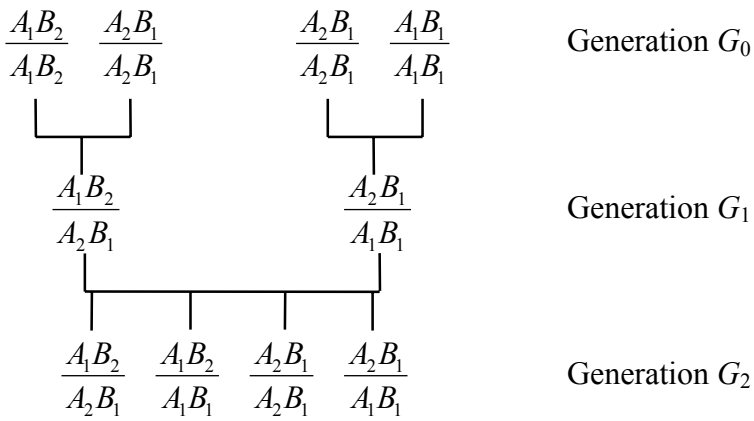
Je mehr Marker und je weniger Allele an den Markern vorliegen, um so höher ist der Anteil der Familien, die bei der Auswertung nicht berücksichtigt werden. Durch das Ausschließen von Familien verringert sich nicht nur die Power des durchgeführten Testes, sondern zusätzlich kommt es zu einer Inflation des Fehlers erster Art. Dieses soll an einem Beispiel veranschaulicht werden.

Beispiel für die Inflation des Fehlers erster Art

Man betrachte die beiden eng gekoppelten, biallelischen Genorte A und B mit θ=0, es finde also keine Rekombination zwischen diesen Genorten statt, in einer Population über drei Generationen: Die Generationen G_0 umfasse zwei getrennte Bevölkerungsgruppen: In der ersten Gruppe gebe es nur Kinder aus

Paarungen, bei denen eine Person den Genotyp $\frac{A_1B_2}{A_1B_2}$ und die andere Person den Genotyp $\frac{A_2B_1}{A_2B_1}$ habe. Analog seien in der zweiten Gruppe die Paarungen zwischen Personen mit den Genotypen $\frac{A_2B_1}{A_2B_1}$ und $\frac{A_1B_1}{A_1B_1}$. Der Genotyp der Kinder (Generation G_1) in der ersten Bevölkerungsgruppe kann dann nur $\frac{A_1B_2}{A_2B_1}$ und der Genotyp der Kinder in der zweiten Bevölkerungsgruppe kann nur $\frac{A_2B_1}{A_1B_1}$ sein. In der Generation G_1 kann eine Paarung einer Person der ersten Bevölkerungsgruppe mit einer Person der zweiten Gruppe zu Kindern mit vier verschiedenen Genotypen führen, nämlich:

$$\frac{A_1B_2}{A_2B_1},\ \frac{A_1B_2}{A_1B_1},\ \frac{A_2B_1}{A_2B_1} \text{ und } \frac{A_2B_1}{A_1B_1}.$$



In der Generation der Großeltern G_0 kommen lediglich homozygote Genorte vor. Deshalb kann aus den gezeigten Paarungen in jeder Bevölkerungsgruppe nur jeweils ein Genotyp in der Elterngeneration G_1 entstehen. Die Kinder dieser beiden Eltern können vier verschiedene Genotypen ausbilden. Jeder dieser vier Genotypen tritt unter Annahme der Nullhypothese von fehlender Assoziation und fehlender Kopplung zwischen den Haplotypen und der Krankheit mit einer Wahrscheinlichkeit von P=0.25 auf. Untersucht man Trios mit Eltern, die einen Genotyp haben wie die Personen aus der Generation G_1 in der oben gezeigten Population, und jeweils einem erkrankten Kind, zeigt sich, dass GENEHUNTER Kinder mit dem Genotyp $\frac{A_1B_2}{A_2B_1}$ oder $\frac{A_2B_1}{A_1B_1}$ aus der Analyse ausschließt. Der Grund hierfür liegt in der nicht eindeutig festzustellenden Übertragung von Haplotypen der Eltern auf die Kinder. Diese Kinder entstehen, wenn ein Elter den Haplotyp A_2B_1 überträgt, der andere Elter aber nicht. Haben in einem Trio aber beide Eltern entweder den Haplotyp A_2B_1 übertragen oder den Haplotyp A_2B_1 nicht übertragen, berücksichtigt GENEHUNTER diese. GENEHUNTER berücksichtigt also nur solche Trios nicht, bei denen die Differenz aus der Anzahl der Haplotypen A_2B_1, die übertragen wurden und der Haplotypen A_2B_1, die nicht übertragen wurden, Null ist (b-c=0) Durch das Ignorieren dieser Familen bleibt der Zähler der McNemar-Testgröße

$$TDT = \frac{(b-c)^2}{b+c}$$

unverändert, der Erwartungswert des Nenners halbiert sich. Um einen korrekten p-Wert zu erhalten, muss man also die Testgröße mit den verdoppelten Quantilen der χ^2-Verteilung vergleichen. Ohne diese Verdopplung erhält man durch GENEHUNTER einen erhöhten Fehler erster Art von

$$\alpha = P(TDT > \frac{\chi^2_{1;0,95}}{2}) = P(TDT > 1.92) = 0.1658.$$

Die gezeigte Stammbaumsituation mag konstruiert und realitätsfern erscheinen, sie entspricht aber der mit dem TDT testbaren Nullhypothese
H_0*: Keine Kopplung oder keine Assoziation liegt vor*.

3.2.2 Transmit

Transmit ist eine Software von Clayton (1999), die man kostenlos über das Internet erhält (http://www.mrc-bsu.cam.ac.uk/pub/methodology/genetics/).

Eingabedatei
Die Eingabedatei enthält die Markerdaten im Format des "pedfile" im Linkage-Programm werden von einem ebenfalls bereitgestellten Programm namens "ped2spl" in ein von Transmit lesbares Format transformiert. Dieses Format unterscheidet sich von dem Linkage-Format (siehe Terwilliger und Ott, 1994) lediglich durch "/"-Zeichen zwischen den Allelangaben jedes Genortes. Transmit schätzt das Ausmaß allelischer Assoziation, auch im Fall nicht eindeutiger Markerhaplotypen.

Die ersten sechs Spalten enthalten Familiennummer, Individuennummer, Nummer des Vaters, Nummer der Mutter, Geschlecht (1 entspricht männlich, 2 entspricht weiblich) und schließlich den Krankheitsstatus (1 bedeutet gesund, 2 bedeutet krank). In den letzen Spalten stehen die Allelinformationen der betrachteten (hier zwei) Genorte.

Testverfahren

Der von Clayton programmierte Test basiert auf der Gesamt-Likelihood, die sich als Produkt zweier Faktoren darstellen lässt:

$$LG = LE \cdot LB$$

$$\Leftrightarrow \text{P(Genotyp der Eltern, Genotyp des Kindes|Kind ist krank)}$$
$$= \text{P(Genotyp der Eltern|Kind ist krank)}$$
$$\text{P(Genotyp des Kindes|Genotyp der Eltern, Kind ist krank)}$$

LG steht für die gesamte Likelihood, *LE* für die Likelihood der Eltern und *LB* ist die auf den Genotyp der Eltern (und den Krankheitsstatus des Kindes) bedingte Likelihood.

Die in der obigen Gleichung auftretenden Wahrscheinlichkeiten werden durch folgende Parameter modelliert:

- β_i: Dieser Parameter entspricht dem Logarithmus der HRR's (Relative Risiken der Haplotypen):

$$\beta_i = \ln(HRR_i)$$

- γ_i: Dieser Parameter ist die multinomiale Logittransformation der Haplotyp-Häufigkeiten h_i:

$$h_i = \frac{\exp(\gamma_i)}{\text{Summe}_i \exp(\gamma_i)}.$$

Liegt keine Assoziation vor, sind die Relativen Risiken gleich Eins. Einen Test auf Assoziation erhält man also, indem man die Nullhypothese

$$H_0: \beta=0$$

mit dem Score-Test oder dem Likelihood-Ratio-Test untersucht. Diese beiden Testverfahren sind asymptotisch äquivalent, in Transmit findet der Score-Test Anwendung. Verwendet man zum Testen die gesamte Likelihood *LG*, also auch den auf dem Populationsmodell für die elterlichen Genotypen basierenden Term *LE*, wird im Allgemeinen das vorgegebene Signifikanzniveau überschritten, weil das wahre Populationsmodell unbekannt ist. Deshalb wird in vielen Tests wie dem TDT von Spielman (1993) für biallelische Marker oder dem erweiterten TDT (ETDT) für multiallelische Marker von Sham und Curtis (1995) lediglich die auf den elterlichen Genotypen bedingte Likelihood verwendet.

Ein Kompromiss zwischen diesen beiden Ansätzen (gesamte Likelihood einerseits und bedingte Likelihood andererseits) wird in der Transmit-Software verwirklicht. Für Trios, in denen die elterlichen Genotypen und die Haplotyp-Übertragungen auf das Kind eindeutig sind, wird die auf den elterlichen Genotypen basierende Likelihood *LE* zum Schätzen des Parameters γ, der die Haplotyp-Häufigkeiten beschreibt, verwendet. Zum Schätzen des Parameters β greift das Programm auf den bedingten Teil der Likelihood zurück. In Trios mit mehrdeutigen elterlichen Genotypen oder Haplotyp-Übertragungen gewichtet man die Score-Funktionen, wie sie bei eindeutiger Datenlage auftreten, mit den verschiedenen, mit den Daten konsistenten, Gesamt-Likelihoods. Durch diesen Ansatz erhält man nach Ansicht von Clayton (1999) ein sehr viel robusteres Verfahren als bei der Verwendung des Ansatzes, bei dem nur die gesamte Likelihood verwendet wird.

Ab Version 2.5 des Transmit-Programms wird ein Bootstrap-Verfahren (Efron und Tibshirani, 1993) angeboten, mit dessen Hilfe man p-Werte für die

durchgeführten Tests erhält und somit entscheiden kann, ob ein signifikantes Testergebnis vorliegt.

Fehlermeldungen

In wenigen Ausnahmefällen kommt es entgegen der Beschreibung durch den Programmautoren noch zu Fehlern. Beispielsweise erhält man bei folgendem Eingabe-Datensatz die Meldung, dass ein Fehler bei der Berechnung der Informationsmatrix aufgetreten ist, was darauf zurückzuführen ist, dass diese Matrix nicht positiv semidefinit ist und somit nicht invertierbar ist:

1	1	0	0	1	1	1/2	1/2
1	2	0	0	2	1	1/1	1/2
1	3	1	2	1	2	1/1	1/2
2	4	0	0	1	1	1/2	1/2
2	5	0	0	2	1	1/1	1/2
2	6	3	4	1	2	1/1	1/2
3	7	0	0	1	1	1/2	1/2
3	8	0	0	2	1	1/1	1/2
3	9	7	8	1	2	1/2	1/2
4	10	0	0	1	1	1/2	1/2
4	11	0	0	2	1	1/1	1/2
4	12	10	11	1	2	1/2	1/2.

In der Regel liefert Transmit aber schnell und unkompliziert das Ergebnis der Untersuchung.

3.3 Multi-Marker-TDT

3.3.1 Modell und Notation

Der im Folgenden vorgestellte Test ist insofern als "Multi-Marker-*TDT*" zu bezeichnen, als er wie der TDT von Spielman (1993)

- familienbasiert ist und somit keine Information über die Populationsstruktur benötigt
- nicht voraussetzt, dass die Markerallel- oder Haplotyp-Häufigkeiten bereits vor der Untersuchung bekannt sind
- die Nullhypothese testet:
  H_0: *Es liegt keine Kopplung oder keine Assoziation vor*

Das Modell, welches für den Multi-Marker-TDT verwendet wird, ist eine Erweiterung des multiallelischen ETDT von Sham und Curtis (1995). Basierend auf dem Quotienten aus einer Likelihood unter der Nullhypothese und der maximierten Likelihood vergleicht man die Testgröße mit dem kritischen Wert einer χ^2-Verteilung. Ist der Wert der Teststatistik größer als der kritische Wert, so erhält man als Testentscheidung die Ablehnung der Nullhypothese und somit den Nachweis einer Assoziation bei vorliegender Kopplung. Wird die Nullhypothese nicht abgelehnt, so entspricht dieses dem Ergebnis:
Es liegt keine Kopplung oder keine Assoziation vor.
Bedingung für die Anwendbarkeit des Testes ist, dass zwischen den betrachteten Markern keine Rekombination stattfindet (θ=0). Wie beim TDT von Spielman et al. (1993) sind Trios (also Familien bestehend aus Vater, Mutter und erkranktem Kind) Grundlage des Multi-Marker-TDT.

Schätzung der Haplotyp-Häufigkeiten

Im ersten Schritt des Testverfahrens werden die Haplotyp-Häufigkeiten durch Maximum-Likelihood-Schätzung ermittelt. Die Schätzung basiert auf den Personen des Stammbaums, deren Eltern nicht im Stammbaum vertreten sind ("founder"). Das Argument für die Zulässigkeit der Haplotyp-Häufigkeitsschätzung aus der Population, in welcher auf Multi-Marker-Assoziation getestet wird, ist, dass unter der Nullhypothese die Verteilung der Haplotyp-Häufigkeiten in den Kranken nicht von der Verteilung der Haplotyp-Häufigkeiten in den gesunden Personen abweicht. Nur unter der Alternativhypothese kann, da der Krankheitsstatus aller Kinder "krank" ist, die Berücksichtigung ihrer Genotypen oder der parentalen Genotypen zu einer verzerrten Schätzung der Haplotyp-Häufigkeiten führen. Liegt beispielsweise ein Krankheitsgenort A mit dem rezessiven Krankheitsallel A_1 und dem dominanten "normalen" Wildtypallel A_2 innerhalb des untersuchten Haplotyps, müssen also zur Krankheitsentstehung an diesem Genort zwei Allele A_1 vorliegen, so würde dies zu einer erhöhten Schätzung der Häufigkeiten solcher Haplotypen führen, in denen dieses Allel A_1 vorkommt.

Die erste Darstellung von Genfrequenzschätzung findet man bei Cepellini et al. (1955) und wurde im Programm EH von Xie und Ott (1993) verwirklicht. Im Folgenden wird die Schätzung der Haplotyp-Häufigkeiten beschrieben.

Die geschätzten Häufigkeiten der Haplotypen $H_1,...,H_L$ seien mit $\hat{h}_1,...,\hat{h}_L$ bezeichnet. Die Idee zur Schätzung ist: Man modelliert die Wahrscheinlichkeit der elterlichen Genotypen durch die zu schätzenden Parameter $h_1,...,h_L$ und maximiert diesen Ausdruck unter der Nebenbedingung $h_1+...+h_L=1$. Eine Schätzung von Parametern durch Maximieren der Likelihood bezeichnet man als "Maximum-Likelihood-Schätzung". Die Modellgleichung ist wegen der Unabhängigkeit der einzelnen Personen ein Produkt von Wahrscheinlichkeiten. Die Wahrscheinlichkeit für den Genotyp einer Person, die an allen betrachteten Genorten homozygot ist, beträgt

$$P(\text{Person hat den Genotyp } \frac{H_i}{H_i}) = h_i^2 .$$

Die Wahrscheinlichkeit, dass eine Person die beiden Haplotypen H_i und H_j, die sich an genau einem Genort unterscheiden, aufweist, beträgt:

$$P(\text{Person hat den Genotyp } \frac{H_i}{H_j}) = 2 \cdot h_i h_j .$$

Schließlich berechnet man die Wahrscheinlichkeit für einen Genotyp einer Person, der mehr als einen heterozygoten Genort umfasst, nach folgendem Ausdruck:

$$\begin{aligned} &\text{P(Person hat einen Genotyp mit mehr als einem heterozygoten Genort)} \\ = \; & 2 \cdot \frac{1}{2^{Het}} \text{Summe}_{Phasen} h_i h_j . \end{aligned}$$

Der Term *Het* bezeichne die Anzahl der heterozygoten Genorte in dem betrachteten Haplotyp. Die verschieden Phasen erhält man durch Vertauschen der beiden Allele an den heterozygoten Genorten. Die Anzahl verschiedener Phasen beträgt 2^{Het}. Die Parameter für die Haplotyp-Häufigkeiten h_i und h_j sind von der Phase abhängig. Auf einen entsprechenden Index wurde zur Vereinfachung der Notation verzichtet.
Diese geschätzten Haplotyp-Häufigkeiten werden in dem folgenden Modell, welches als Grundlage für den Assoziationstest dient, in Familien mit unbekannter Phase benötigt.

Likelihood-Beitrag bei eindeutiger Phase

Ist in einem Trio die Phase eindeutig und somit bekannt, welche beiden Haplotypen der Eltern auf das Kind übertragen wurden und welche beiden Haplotypen nicht übertragen wurden, sieht der Likelihood-Beitrag dieser Familie wie folgt aus:

$$L_i = \frac{\exp(\lambda_j - \lambda_{j'})}{1+\exp(\lambda_j - \lambda_{j'})} \cdot \frac{\exp(\lambda_k - \lambda_{k'})}{1+\exp(\lambda_k - \lambda_{k'})}. \qquad (3.1)$$

Der Index i bezeichnet dabei das i-te Trio, j und k sind die Nummern der Haplotypen, die vom Vater und von der Mutter auf das erkrankte Kind übertragen wurden und j' und k' sind die Nummern der nicht-übertragenen Haplotypen. Die Haplotyp-Indices haben folgende Bedeutung:

Index	Haplotyp
1	A_1B_1
2	A_1B_2
3	A_2B_1
4	A_2B_2

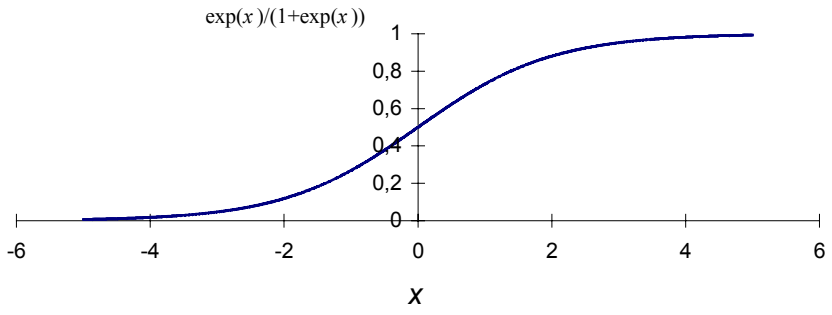


Abbildung 3.5: Likelihoodbeitrag eines Elternteils in Abhängigkeit von der Differenz λ_j-$\lambda_{j'}$:=x, unter H_0 gilt: x=0.

Wenn die Phase bei zwei eng gekoppelten biallelischen Markern eindeutig ist, bestehen die Likelihoodbeiträge jedes Elternteils aus einem Bruch, der die Form eines Faktors der Gleichung (3.1) aufweist. Sind übertragener und nicht übertragener Haplotyp identisch, liefert der Elter keine für die Untersuchung auf Assoziation relevante Information (siehe Tabelle 3.1). Dann werden in der Testgröße Zähler und Nenner mit dem Faktor

$$\frac{\exp(\lambda_j - \lambda_{j'})}{1+\exp(\lambda_j - \lambda_{j'})} = \frac{1}{2}, \text{ wobei } j = j',$$

multipliziert. Die Testgröße ändert ihren Wert also nicht.

Tabelle 3.1: Likelihoodbeitrag eines Eltern bei zwei biallelischen Markern und eindeutiger Phase

		Nicht übertragener Haplotyp			
		A_1B_1	A_1B_2	A_2B_1	A_2B_2
Übertragener Haplotyp	A_1B_1	$\frac{\exp(\lambda_1-\lambda_1)}{1+\exp(\lambda_1-\lambda_1)}$	$\frac{\exp(\lambda_1-\lambda_2)}{1+\exp(\lambda_1-\lambda_2)}$	$\frac{\exp(\lambda_1-\lambda_3)}{1+\exp(\lambda_1-\lambda_3)}$	$\frac{\exp(\lambda_1-\lambda_4)}{1+\exp(\lambda_1-\lambda_4)}$
	A_1B_2	$\frac{\exp(\lambda_2-\lambda_1)}{1+\exp(\lambda_2-\lambda_1)}$	$\frac{\exp(\lambda_2-\lambda_2)}{1+\exp(\lambda_2-\lambda_2)}$	$\frac{\exp(\lambda_1-\lambda_2)}{1+\exp(\lambda_1-\lambda_2)}$	$\frac{\exp(\lambda_2-\lambda_4)}{1+\exp(\lambda_2-\lambda_4)}$
	A_2B_1	$\frac{\exp(\lambda_3-\lambda_1)}{1+\exp(\lambda_3-\lambda_1)}$	$\frac{\exp(\lambda_3-\lambda_2)}{1+\exp(\lambda_3-\lambda_2)}$	$\frac{\exp(\lambda_3-\lambda_3)}{1+\exp(\lambda_3-\lambda_3)}$	$\frac{\exp(\lambda_3-\lambda_4)}{1+\exp(\lambda_3-\lambda_4)}$
	A_2B_2	$\frac{\exp(\lambda_4-\lambda_1)}{1+\exp(\lambda_4-\lambda_1)}$	$\frac{\exp(\lambda_4-\lambda_2)}{1+\exp(\lambda_4-\lambda_2)}$	$\frac{\exp(\lambda_4-\lambda_3)}{1+\exp(\lambda_4-\lambda_3)}$	$\frac{\exp(\lambda_4-\lambda_4)}{1+\exp(\lambda_4-\lambda_4)}$

Likelihood-Beitrag bei mehrdeutiger Phase

Ist in einem Trio die Phase nicht eindeutig, besteht sein Likelihood-Beitrag aus einer Summe von Termen wie in Gleichung (3.1). Summiert wird über die möglichen Phasenerklärungen, wobei jeder Summand einen Gewichtungsfaktor erhält, den man mittels der geschätzten Haplotyp-Häufigkeiten erhält. Der Index *r* bezeichnet die beiden Eltern jedes Kindes und die Indices *T* und *NT* geben an, ob es sich um den übertragenen Haplotyp (*T* wie "transmitted") oder um den nicht übertragenen Haplotyp (*NT* wie "not transmitted") handelt.

$$L_i = \text{Summe}_{Phasen} \frac{\prod_{r=1}^{2} \hat{h}_{r,T} \hat{h}_{r,NT}}{\text{Summe}_{Phasen} \prod_{r=1}^{2} \hat{h}_{r,T} \hat{h}_{r,NT}} \cdot \frac{\exp(\lambda_j - \lambda_{j'})}{1+\exp(\lambda_j - \lambda_{j'})} \cdot \frac{\exp(\lambda_k - \lambda_{k'})}{1+\exp(\lambda_k - \lambda_{k'})}$$

.

Beispiel (Phase mehrdeutig)

Man betrachte zum Beispiel das folgende Trio mit zwei biallelischen Markern: Am Genort *A* mögen beide Eltern und das Kind (gleichermaßen) den heterozygoten Genotyp (A_1,A_2) aufweisen, am Genort *B* seien der erste Elter und das Kind heterozygot (B_1,B_2) und der zweite Elter homozygot (B_1,B_1). Die Haplotypen seien nicht bekannt.

$$\frac{A_1}{A_2}\ \frac{B_1}{B_2} \times \frac{A_1}{A_2}\ \frac{B_1}{B_1} \rightarrow \frac{A_1}{A_2}\ \frac{B_1}{B_2}. \qquad (3.2)$$

Die aus den Daten geschätzten Haplotyp-Häufigkeiten der Haplotypen A_1B_1; A_1B_2; A_2B_1 und A_2B_2 seien

$$\hat{h}_{11} = 0.4,\ \hat{h}_{12} = 0.3,\ \hat{h}_{21} = 0.2,\ \hat{h}_{22} = 0.1\,.$$

Für obiges Trio ist die Phase nicht eindeutig, da die vorliegende Konstellation mit zwei verschiedenen Haplotyp-Erklärungen vereinbar ist. Es gibt die folgenden zwei Möglichkeiten (die übertragenen Haplotypen stehen bei den Eltern jeweils über dem Bruchstrich):

$$1)\quad \frac{A_1B_2}{A_2B_1} \times \frac{A_2B_1}{A_1B_1} \rightarrow \frac{A_1B_2}{A_2B_1}$$

$$2)\quad \frac{A_2B_2}{A_1B_1} \times \frac{A_1B_1}{A_2B_1} \rightarrow \frac{A_2B_2}{A_1B_1}$$

Bei der ersten Möglichkeit werden die Haplotypen A_1B_2 und A_2B_1 übertragen, die Haplotypen A_2B_1 und A_1B_1 aber nicht. Dieses geschieht mit einer Wahrscheinlichkeit von

$$P(Phase = 1)$$

$$= \frac{2\hat{h}_{12}\hat{h}_{21} \cdot 2\hat{h}_{21}\hat{h}_{11}}{2\hat{h}_{12}\hat{h}_{21} \cdot 2\hat{h}_{21}\hat{h}_{11} + 2\hat{h}_{22}\hat{h}_{11} \cdot 2\hat{h}_{11}\hat{h}_{21}}$$

$$= \frac{2 \cdot 0.3 \cdot 0.2 \cdot 2 \cdot 0.2 \cdot 0.4}{2 \cdot 0.3 \cdot 0.2 \cdot 2 \cdot 0.2 \cdot 0.4 + 2 \cdot 0.1 \cdot 0.4 \cdot 2 \cdot 0.4 \cdot 0.2}$$

$$= 0.6.$$

Entsprechend gilt für die Wahrscheinlichkeit der zweiten Möglichkeit

$$P(Phase = 2)$$
$$= 1 - P(Phase = 1)$$
$$= 0.4.$$

Somit gilt für den Likelihood-Beitrag dieses Trios:

$$L_i = 0.6 \cdot \frac{\exp(\lambda_2 - \lambda_3)}{1+\exp(\lambda_2 - \lambda_3)} \cdot \frac{\exp(\lambda_3 - \lambda_1)}{1+\exp(\lambda_3 - \lambda_1)} + 0.4 \cdot \frac{\exp(\lambda_4 - \lambda_1)}{1+\exp(\lambda_4 - \lambda_1)} \cdot \frac{\exp(\lambda_1 - \lambda_3)}{1+\exp(\lambda_1 - \lambda_3)}.$$

Bestehen die Daten nur aus Trios mit elterlichen Genotypen wie im Trio (3.2), lautet die Schätzung der Haplotyp-Häufigkeiten

$$\hat{h}_{11} = 0.375,\ \hat{h}_{12} = 0.125,\ \hat{h}_{21} = 0.375,\ \hat{h}_{22} = 0.125.$$

Die Wahrscheinlichkeit für Phase 1 ist dann gleich der Wahrscheinlichkeit für Phase 2:

$$P(\text{Phase} = 1)$$

$$= \frac{2\hat{h}_{12}\hat{h}_{21} \cdot 2\hat{h}_{21}\hat{h}_{11}}{2\hat{h}_{12}\hat{h}_{21} \cdot 2\hat{h}_{21}\hat{h}_{11} + 2\hat{h}_{22}\hat{h}_{11} \cdot 2\hat{h}_{11}\hat{h}_{21}}$$

$$= \frac{2 \cdot 0.125 \cdot 0.375 \cdot 2 \cdot 0.375 \cdot 0.375}{2 \cdot 0.125 \cdot 0.375 \cdot 2 \cdot 0.375 \cdot 0.375 + 2 \cdot 0.125 \cdot 0.375 \cdot 2 \cdot 0.375 \cdot 0.375}$$

$$= 0.5$$

$$= P(\text{Phase} = 2),$$

Man bestimmt also alle Möglichkeiten, die elterlichen Allele so zu Haplotypen zusammenzusetzten, dass der Genotyp des Kindes entstehen kann und gewichtet die verschiedenen Möglichkeiten analog zu den geschätzten Haplotyp-Häufigkeiten.

Beispiel (Phase eindeutig)

Die Phase wird eindeutig, wenn einer der Eltern oder das Kind statt des heterozygoten Genotyps einen homozygoten Genotyp (A_1,A_1) oder (A_2,A_2) vorliegen hat.

$$\frac{A_1B_2}{A_2B_1} \times \frac{A_1B_1}{A_2B_1} \rightarrow \frac{A_1B_2}{A_1B_1}.$$

Die Likelihood für dieses Trio lautet zum Beispiel:

$$L_i = \frac{\exp(\lambda_2 - \lambda_3)}{1+\exp(\lambda_2 - \lambda_3)} \cdot \frac{\exp(\lambda_1 - \lambda_3)}{1+\exp(\lambda_1 - \lambda_3)}.$$

Bestimmung der Testgröße und ihrer Verteilung

Liegen m verschiedene Haplotypen vor, lautet die zu testende Nullhypothese:

H_0: $\lambda_1 = \lambda_2 = ... = \lambda_m$.

Diese wird getestet gegen die Alternativhypothese, dass nicht alle Parameter $\lambda_1, \lambda_2 ..., \lambda_m$ den gleichen Wert haben. Die Testgröße des Likelihood-Ratio-Testes (*LRT*) lautet:

$$LRT = -2 \ln \frac{L(B_0)}{L(B)},$$

wobei $L(B_0)$ die Likelihood unter der Nullhypothese und $L(B)$ die über den Parameterraum maximierte Likelihood ist. Gemäß der allgemeinen Testtheorie (Rao, 1973) ist diese Größe χ^2-verteilt mit (m-1) Freiheitsgraden.

Ein SAS-Programm (SAS Institute Inc., 1990), welches den Multi-Marker-TDT für alle vollständig typisierten Trios durchführt und zusätzlich zu den p-Werten

die Größe der Teststatistik und die Anzahl der Freiheitsgrade ausgibt, wurde erstellt und ist vom Autor zu beziehen.

3.3.2 Eine Anwendung auf simulierte Daten

Die simulierten Daten, an denen das Multi-Marker-Programm getestet wurde, entstammen einem Datensatz des Genetic Analysis Workshop 9 (GAW9) (Hodge, 1995).

Beschreibung der Daten

Dieser Datensatz umfasst 200 Kernfamilien (also Eltern mit ihren Kindern) mit mindestens einem erkrankten Kind und 100 Kernfamilien, die nur aus gesunden Personen bestehen. Um zu Trios zu gelangen wurde in größeren Familien alle Kinder bis auf das erste erkrankte ignoriert. Die Daten wurden derart simuliert dass vier Gene auf verschiedenen Chromosomen an der Krankheitsentstehung beteiligt sind. Von den typisierten Markern sind zwei mit jeweils einem der insgesamt vier Krankheitsgenorte identisch. Diese beiden Genorte liegen auf Chromosom 1 und 5 und haben jeweils acht Allele. In folgender Weise lässt sich ein solcher Genort in drei neue biallelische Genorte aufteilen (siehe Tabelle 3.2).

Tabelle 3.2: Aufspaltung eines Genortes mit acht Allelen auf drei biallelische Genorte

Allel am ursprünglichen Genort	Allel am ersten biallelischen Genort	Allel am zweiten biallelischen Genort	Allel am dritten biallelischen Genort
1	1	1	1
2	1	1	2
3	1	2	1
4	1	2	2
5	2	1	1
6	2	1	2
7	2	2	1
8	2	2	2

Ergebnisse der Auswertung

Nun "vergisst" man, dass es sich bei den Genorten um Krankheitsgenorte handelt und verfährt so, als seien sie Marker. Nach der Aufteilung auf drei Genorte kann man den Multi-Marker-TDT auf einen, zwei und alle drei Genorte anwenden. Dabei erhält man die folgenden Ergebnisse.

Tabelle 3.3: p-Werte auf Chromosom 1

Gen-ort	Länge des Haplotyps		
	1	2	3
1	0.40		
		0.12	
2	0.063		0.0017
		0.21	
3	0.88		

Tabelle 3.4: p-Werte auf Chromosom 5

Gen-ort	Länge des Haplotyps		
	1	2	3
1	$4.19 \cdot 10^{-6}$	$6.23 \cdot 10^{-9}$	
2	$9.20 \cdot 10^{-7}$	$3.36 \cdot 10^{-6}$	$1.37 \cdot 10^{-10}$
3	$4.3 \cdot 10^{-5}$		

Die Tabellen 3.3 und 3.4 stellen die p-Werte der Multi-Marker-Testgröße auf den Chromosomen eins und fünf dar. Zum einen fällt auf, dass die p-Werte, die man bei der Untersuchung des Chromosoms eins erhält, im Vergleich zu den Werten von Chromosom fünf relativ groß sind. Dies entspricht der Tatsache, dass die Assoziation auf Chromosom fünf deulich stärker ist als auf Chromosom eins Zum anderen ist in beiden Tabellen ein Abfall der p-Werte von links nach rechts, also in Richtung der längeren Haplotypen vorhanden. Auf Chromosom fünf wird die vorhandene Assoziation, unabhängig von der Länge des Haplotyps, gefunden, weil alle dort alle p-Werte kleiner als 0.5 sind. Die Assoziation auf dem Chromosom eins ist aber durch die Analyse mit einem oder zwei Genorten nicht nachweisbar, sondern wird nur bei Verwendung des Haplotyps über die volle Länge von drei Markern gefunden.

3.3.3 Eine Anwendung auf reale Daten

Schließlich wurde mit dem Multi-Marker-TDT auch ein realer Datensatz mit Patienten, die von der Krankheit Psoriasis vulgaris betroffen sind, untersucht. Psoriasis vulgaris oder Schuppenflechte ist eine häufige, (etwa 2% der

europäischen Bevölkerung sind betroffen,) genetisch komplexe Hauterkrankung. Verursacht wird die Psoriasis nach heutiger Ansicht sowohl durch genetische als auch durch Umweltfaktoren (z.B. Reizung der Haut oder Stress). Bei Ausbruch der Krankheit kommt es zu einer stark beschleunigten Epidermisbildung, was die Entstehung einer Schuppenschicht auf der Hautoberfläche zur Folge hat. Die Therapie der Psoriasis stellt immer noch ein Problem dar. In Frage kommen Salben, Licht- und Badebehandlung und auch eine innere Behandlung (Tabletten), was für die meisten Patienten zumindest zu einer Linderung der Symptome führt (Bhalerao und Bowcock, 1998).

Beschreibung des Datensatzes

Die untersuchten Daten umfassen 52 Trios, die in Berlin und Münster sowie in der Umgebung dieser beiden Städte erhoben wurden. Typisiert wurde das Kollektiv an sechs eng benachbarten Genorten auf dem kurzen Arm des Chromosoms sechs (6p21.3). Zwei der typisierten Genorte kommen aus der HLA (human leucozyte antigen)-Region (HLA- B und HLA-C). Diese beiden Genorte sind hochpolymorph: 28 Allele am Marker HLA-B und 15 Allele am Marker HLA-C. Die übrigen vier Genorte liegen im Bereich des Corneodesmosin-Gens, welches 160 Kb telomerisch zum HLA-C liegt. Da der dritte von diesen vier Genorten keine Variation in dem Kollektiv (alle typisierten Personen haben an dieser Position 1240 die Base Guanin) zeigt und somit weder für Kopplung noch für Assoziation informativ ist, wurde dieser nicht für die Auswertung verwendet. Die drei Corneodesmosin-Marker, welche Variation in den typisierten Personen zeigen, sind allesamt biallelisch, also SNPs. Die Corneodesmosin-Marker mit den Basenpositionen 619 und 1243 tragen die Basen Cytosin oder Thymin, der Corneodesmosin-Marker auf den Positionen 1236 trägt die Basen Guanin oder Thymin. Die angegebenen Positionen entsprechen denen aus der Veröffentlichung von Ishihara et al.

(1996). Die fünf Marker mit Variation seien mit *M*1 bis *M*5 bezeichnet (siehe Tabelle 3.5).

Tabelle 3.5: Verwendete Marker und deren Bezeichnung

Typisierter Marker (Basenposition)	Bezeichnung	Anzahl der Allele
HLA-B	*M*1	28
HLA-C	*M*2	15
erster Corneodesmosin-Marker (619)	*M*3	2
zweiter Corneodesmosin-Marker (1236)	*M*4	2
dritter Corneodesmosin-Marker (1240)	_____	1
vierter Corneodesmosin-Marker (1243)	*M*5	2

Die hohe Anzahl von Allelen bei den Markern *M*1 und *M*2 ermöglicht es, in allen Trios die Haplotypen, die die ersten beiden Marker umfassen, eindeutig zu bestimmen. Die zusätzliche Verwendung der Corneodesmosin-Marker, die jeweils nur zwei Allele haben, erlaubt dagegen nicht die eindeutige Bestimmung der Haplotypen über die volle Länge der fünf Marker hinweg. Deshalb werden die Haplotyp-Häufigkeiten mit Hilfe der elterlichen Genotypinformation geschätzt um den Assoziationstest anzuschließen.

Ergebnisse

Mit Hilfe eines SAS-Programms, welches das in Kapitel 3.3.1 beschriebene Testverfahren anwendet, wurden die in Tabelle 3.6 zusammengefassten Ergebnisse erhalten. Insgesamt wurden 15 Tests durchgeführt. Die erste Spalte der Tabelle enthält eine fortlaufende Testnummer. In Spalte zwei steht die Länge des Haplotyps, der jeweils auf signifikante Assoziation untersucht wurde. Spalte drei zeigt die Marker, die den getesteten Haplotyp bilden. In der letzten Spalte findet man den p-Wert des jeweiligen Testes.

Tabelle 3.6: Auswertung der Psoriasis-Daten

Testnummer	Länge des Haplotyps	Verwendete Marker	Trioanzahl	p-Wert
1	5	*M*1 *M*2 *M*3 *M*4 *M*5	43	0.7169
2	4	*M*1 *M*2 *M*3 *M*4	43	0.3717
3	4	*M*2 *M*3 *M*4 *M*5	43	0.1084
4	3	*M*1 *M*2 *M*3	44	0.1084
5	3	*M*2 *M*3 *M*4	43	0.0055
6	3	*M*3 *M*4 *M*5	43	0.0142
7	2	*M*1 *M*2	52	0.0026
8	2	*M*2 *M*3	44	0.0047
9	2	*M*3 *M*4	43	0.9482
10	2	*M*4 *M*5	46	0.0078
11	1	*M*1	52	0.0071
12	1	*M*2	52	0.0038
13	1	*M*3	44	0.8474
14	1	*M*4	46	0.3345
15	1	*M*5	47	0.0025

Die Tabelle 3.6 fasst die Ergebnisse der mit den Markern aus der HLA- und denen aus der Corneodesmosin-Region durchgeführten Tests zusammen. Die Signifikanzgrenze von α=0.05 muss, um für multiples Testen zu korrigieren, nach Bonferroni (Horn und Vollandt, 1995) durch die Anzahl der durchgeführten Tests geteilt werden. Da die Anzahl der durchgeführten Tests 15 beträgt, sind nur solche Tests signifikant, bei denen der p-Wert einen Wert von $\frac{0.05}{15} = 0.00\overline{3}$ unterschreitet. Signifikant sind also Testergebnis sieben und 15. Die Signifikanz am Marker M5 lässt sich mittels des Multi-Marker-TDT, also

mit Haplotypen der Länge zwei bis fünf nicht bestätigen, was zumindest teilweise auf den geringen Stichprobenumfang zurückzuführen ist. Die zweite Signifikanz ergibt sich in der HLA-Region, wodurch der bereits veröffentlichte Assoziationsbefund von Schmitt-Egenolf (1999) bestätigt wird. In dieser Veröffentlichung wird eine positive Assoziation zwischen Psoriasis und dem Haplotyp bestehend aus dem Allel 5701 am Marker *M*1 (HLA-B) und dem Allel 0602_95T (dieses Allel hat auch die Bezeichnung Cw-6) am Marker *M*2 (HLA-C) nachgewiesen. Deshalb wurde der Haplotyp, bestehend aus dem Allel 5701 am HLA-B und 0602_95T getestet. Man fasst also alle Allele außer Allel 5701 am Marker *M*1 in eine Gruppe zusammen und bezeichnet sie als "Allel 1 des ersten Markers" und fasst ebenso am Marker *M*2 alle Allele außer Allel 0602_95T zu einer Gruppe "Allel 1 des zweiten Markers" zusammen. Diese beiden "biallelischen" Marker führen bei der Auswertung mit dem SAS-Programm zu einem hochsignifikanten p-Wert von 0.00000738. Ohne die benutzte Vorinformation über Kandidatenallele hätte man alle möglichen Allelkombinationen auf diesen beiden Allelen auf Signifikanz testen müssen. Die Anzahl solcher Tests liegt bei

$$\begin{aligned}&\text{Anzahl der Allele an } M1 \cdot \text{Anzahl der Allele an } M2\\ &= 28 \cdot 15\\ &= 420.\end{aligned}$$

Durch eine entsprechende Bonferroni-Korrektur ($\alpha_{korrigiert}$=0.05/420=0.00012) zeigt sich, dass die Signifikanz auch bei Berücksichtigung des multiplen Testens (Horn und Vollandt, 1995) bestehen bleibt.

Kapitel 4

Zusammenfassung und Ausblick

In der vorliegenden Arbeit werden statistische Methoden zur familienbasierten Assoziationsanalyse behandelt. Für diese Methoden typisiert man in der Regel Kernfamilien, also jeweils ein erkranktes Kind mit seinen beiden Eltern. Möchte man dann auf Assoziation zwischen der Krankheit und den Allelen am Marker prüfen, verwendet man, falls der Marker biallelisch ist, den TDT (Spielman et al., 1993) und, falls der Marker multiallelisch ist, den ETDT (Sham und Curtis, 1995).

Bei Krankheiten, die sich erst im hohen Alter manifestieren, ist es schwierig, ausreichend viele Trios zu typisieren, weil dann die Eltern des Erkrankten häufig nicht mehr leben und somit die parentalen Genotypen nicht bestimmt werden können. Um die fehlende Information zu "ersetzen", sollte man dann die Genotypen von Verwandten der erkrankten Person, gut geeignet sind hierfür die Geschwister, bestimmen. Danach kann man die Allelhäufigkeiten der gesunden Geschwister mit den Allelhäufigkeiten der kranken Geschwister vergleichen. Alternativ ist es bei bestimmten Genotypkonstellationen der Geschwister möglich, die fehlenden, parentalen Genotypen zu rekonstruieren und mit Hilfe dieser einen Test auf Assoziation durchzuführen.

Methoden zu familienbasierten Assoziationsanalysen bei mehr als einem Marker werden zwar von den Programmen GENEHUNTER und Transmit angeboten, diese sind aber noch verbesserungsbedürftig.

Aufbauend auf dem ETDT-Ansatz (Sham und Curtis, 1995) wird im Kapitel 3.3 eine neue Methode für die Durchführung einer Assoziationsanalyse mit mehr

als einem Marker vorgestellt. Schließlich wird diese Methode auf einen simulierten Datensatz des Genetic Analysis Workshop 9 und auf einen realen Datensatz angewendet. Die Ergebnisse deuten darauf hin, dass die Multi-Marker-Analyse insbesondere in Situationen mit SNPs der Analyse mit Methoden für nur einen Marker vorzuziehen ist.

Entsprechend den Erweiterungen des TDT mit einem Marker sollte sich auch der Multi-Marker-TDT auf Situationen ohne Eltern erweitern lassen. Mit diesem Ansatz könnte man dann Daten mit verschiedenen Anzahlen von Markern und auch mit verschiedenen Familientypen auswerten. Möglicherweise wird man auf diese Art weitere bisher unbekannte Krankheitsgenorte entdecken und somit zur Behandlung dieser Krankheiten einen ersten Schritt machen können.

Literaturverzeichnis

Bhalerao J, Bowcock AM (1998): *The genetics of psoriasis: a complex disorder of the skin and immune system*. Human Molecular Genetics **7(10)**: 1537-1545.

Cavalli-Sforza LL, Bodmer WF (1971): *The Genetics of Human Populations.* W. H. Freeman & Co., San Francisco, CA.

Cepellini R, Siniscalco M, Smith CAB (1955): *The estimation of gene frequencies in a random-mating population*. Annals of Human Genetics **20**: 97

Collins FS (1999): *Shattuck Lecture – Medical and societal consequences of the Human Genome Project.* New England Journal of Medicine **341**: 28-37.

Efron B, Tibshirani RJ (1993): *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Hodge SE (1995): *An oligogenic disease displaying weak marker associations: A summary of contributions to problem 1 of GAW9*. Genetic Epidemiology **12**: 545-554.

Hodge S, Boehnke M, Spence MA (1999): *Loss of information due to ambiguous haplotyping of SNPs.* Nature Genetics **21:** 360-361.

Horn M, Vollandt R (1995): *Multiple Tests und Auswahlverfahren*. Stuttgart; Jena; New York: G. Fischer.

Horvath S, Laird NM (1998*)*: *A discordant-sibship test for disequilibrium and linkage: No need for parental data*. American Journal of Human Genetics **63**: 1886-1897.

Ishihara M, Yamagata N, Ohno S, Naruse T, Ando A, Kwata H, Ozawa A, Ohkido M, Mizuki N, Shiina T, Ando H, Inoko H (1996): *Genetic polymorphisms in the keratin-like S gene within the human major histocompatibility complex and association analysis on the susceptibility to psoriasis vulgaris*. Tissue antigens **48**: 182-186.

Knapp M (1999): *The Transmission/Disequilibrium Test and parental genotype reconstruction: The reconstruction-combined Transmission/Disequilibrium Test*. American Journal of Human Genetics **64**: 861-870.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996): *Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach*. American Journal of Human Genetics **58**:1347-1363.

Lazzeroni LL, Lange K (1998): *A conditional inference framework for extending the Transmission/Disequilibrium Test*. Human Heredity **48**: 67-81.

McNemar Q (1947): *Note on sampling error of the differences between correlated proportions or percentages*. Psychometrika **12 (153)**: 157.

Müller PH (1991): *Lexikon der Stochastik: Wahrscheinlichkeitsrechnung und Mathematische Statistik*, Berlin: 5. Auflage, Akademie Verlag.

Rao CR (1973): *Linear statistical inference and its applications*, New York, Wiley.

Risch N, Merikangas K (1996): *The future of genetic studies of complex human diseases*. Science **268**: 1516-1517.

SAS Institute Inc. (1990): *SAS language: reference*, version 6, 1st ed. SAS Institute, Cary, NC.

Schmitt-Egenolf M, Windemuth C, Albis-Camps M, von Engelhard B, Sterry W, Traupe H, Blasczyk R (1999): *Transmission disequilibrium of HLA-Cw 0602 and HLA-B 5701 in psoriasis suggests a susceptibility locus between HLA-C and HLA-B.* Journal of Investigative Dermatology **113:** 396.

Sham PC, Curtis D (1995): *An extended transmission/disequilibrium test (TDT) for multi-allele marker loci*. American Journal of Human Genetics **59**: 323-336.

Spielman RS, McGinnis RE, Ewens WJ (1993): *Transmission test for linkage disequilibrium: the insulin-dependent diabetes mellitus (IDDM).* American Journal of Human Genetics **52**: 506-516.

Spielman RS, Ewens WJ (1998): *A sibship test for linkage in the presence of association: The sib Transmission/Disequilibrium Test*. American Journal of Human Genetics **62**: 450-458.

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

Wilson SR (1997): *On extending the transmission/disequilibrium test (TDT)*. Annals of Human Genetics, **61(2)**, 151-161.

Xie X, Ott J (1993): *Testing linkage disequilibrium between a disease gene and marker loci.* American Journal of Human Genetics, suppl **53**, 1107.

Herrn Prof. Dr. Max P. Baur danke ich für die Schaffung der Rahmenbedingung, die die Entstehung dieser Arbeit ermöglichten.

Herrn Prof. Dr. Wolfgang Alt danke ich für die wertvollen Hinweise und die freundliche Unterstützung.

Herrn PD Dr. Michael Knapp danke ich für die Überlassung des Themas und seine stete Bereitschaft zur Klärung auftauchender Probleme.

Frau Dr. Christine Windemuth-Kieselbach und Herrrn Dr. Stefan Horvath danke ich für ihre Anregungen und die kollegiale Zusammenarbeit.

Herrn Prof. Dr. Marcus Schmitt-Egenolf danke ich für die Erlaubnis, die Daten des Chromosoms 6 auszuwerten.

Der Deutschen Forschungsgemeinschaft danke ich für ihre finanzielle Unterstützung im Rahmen des Graduiertenkollegs 246 "Pathogenese von Krankheiten des Nervensystems".