Kapitel 2

Methoden

Das statistische Vorgehen hier ist durch eine Reihe von Besonderheiten gegenüber bewährten biometrisch-genetischen Modellen gekennzeichnet, die durch spezielle Fragestellungen und das über viele Jahre realisierte Versuchsdesign der zugrunde liegenden Studie bedingt sind.

Zunächst wird die Studienpopulation beschrieben und gemessene Phänotypen und Genotypen definiert. Die Analysestrategie motiviert die methodische Herangehensweise. Statistische Modelle, auf denen die Analyse aufbaut, werden eingeführt, wobei der Schwerpunkt auf der Modellierung familienbasierter Assoziation liegt.

Um zwischen den theoretischen Parametern der Modelle (bezogen auf die Population) und den Schätzern (bezogen auf eine Stichprobe) unterscheiden zu können, werden für die Parameter griechische und für die Schätzer lateinische Buchstaben verwendet.

2.1 Studienpopulation

Aus einem Probandenkollektiv, das im Rahmen einer prospektiven Kohortenstudie 1995-1999 an der Franz-Volhard Klinik durch die „Genetic Fieldworking Unit“ rekrutiert wurde (Schuster et al., 1998), wurden 219 Großfamilien aus Deutschland ausgewählt. Der Begriff „Großfamilien“ beeinhaltet sowohl Kernfamilien (Eltern und Kinder), als auch Familien, die mehrere Generationen repräsentieren (Tanten, Onkel, Nichten, Neffen, Großeltern, etc.) Die Rekrutierung der Probanden erfolgte über einen Patienten mit einer koronaren Herz-Erkrankung. Ausgehend von diesen Index-Patienten wurden Verwandte zweiten und höheren Grades sowie Verwandte des (Ehe-) Partners ausgewählt. Die Index-Patienten selbst wurden nicht in die Analyse einbezogen. Probanden, bei denen eine familiäre Lipidstoffwechselstörung, eine bekannte sekundäre Hyperlipidämie oder eine bekannte klinisch manifeste Arteriosklerose vorlag, wurden von der Analyse ausgeschlossen. Die erhaltene Stichprobe umfasst 1054 Patienten, davon 514 Männer und 540 Frauen in 188 Familien mit 3 bis 5 Personen, 17 Familien mit 6 bis 9 Personen, 14 Familien mit 10 bis 14 Personen.

Für eine genetisch-epidemiologische Vergleichsstudie, die den zweiten Teil der vorliegenden Arbeit darstellt, wurde eine unabhängige Stichprobe aus der Schweiz herangezogen.

Die Schweizer populationsbasierte Studie (Erhebung 1999-2000) repräsentiert französisch sprechende Erwachsene (35-74 Jahre) aus dem Kanton Genf. Die Stichprobe umfasst 1708 unverwandte Personen, 846 Männer und 862 Frauen, ohne Anzeichen von Herzkreislauferkrankungen (Morabia et al., 2003b).

Die zugrunde liegende Population beider Studien ist vergleichbar. Es handelt sich jeweils um Individuen mit Cholesterinkonzentrationen im normalen Bereich (ohne Medikation). Auch die Hypothese ist ähnlich: Nachweis von Assoziation häufiger genetischer Varianten in Kandidatengenen des Lipoproteinmetabolismus. Das Studiendesign unterscheidet sich, d.h. familienbasierte versus unverwandte Stichprobe, quantitativ versus qualitativer Phänotyp. Zur Durchführung der Vergleichsstudie wurde das Studiendesign vereinheitlicht (siehe unten).

2.2 Definition und Bestimmung der Phänotypen

Von allen Probanden beider Kohorten wurden die Serumlipidwerte Gesamtcholesterin (TC), LDL-Cholesterin (LDL), HDL-Cholesterin (HDL) und Triglyzeride (TG) bestimmt (Kapitel 3, Tabelle 3.1).

Zur Analyse der quantitativen Lipidphänotypen in der Berliner Familienstudie wurden die Daten logarithmisch transformiert, so dass sie approximativ normalverteilt sind. Individuen mit Cholesterinkonzentrationen im Risikobereich der Verteilungen wurden ausgeschlossen.

Die Verteilungsgüte der transformierten Lipidwerte wurde in einem Histogramm geprüft (Kapitel 3, Abbildung 3.1). Zusätzlich ist die Dichte der Standardnormalverteilung angegeben. Die Anpassung hängt von Schiefe (skewness) und Wölbung (kurtosis) der empirischen Verteilung ab. Im Fall der theoretischen Normalverteilung haben beide Kennzahlen den Wert 0. Die Kolmogorov-Smirnov-Statistik (D) basiert auf der größten vertikalen Differenz zwischen der theoretischen Verteilungsfunktion $F(x)$ und der empirischen Verteilungsfunktion $F_n(x)$ und wird als p-Wert aus dem dazugehörigen Test angegeben.

Die Genfer Studie war als qualitative Studie angelegt. Für die Vergleichsanalyse wurden entsprechend dem Genfer Studiendesign (Morabia et al., 2003b) in den unabhängigen Stichproben separat phänotypisch unterschiedliche Individuen („Fälle“ und „Kontrollen“) selektiert. Diese wurden entsprechend ihrem höchsten bzw. niedrigsten Tertil (T1=33.3%- und T3=66.7%-Perzentil) der geschlechtsspezifischen LDL- bzw. HDL-Verteilung klassifiziert. Daraus resultiert ein kombinierter qualitativer Lipid-Phänotyp. Um einen altersbedingten Selektionsfehler in der Berliner Fall-Kontroll-Auswahl zu vermeiden, wurden die Lipidwerte

vor der Klassifikation für das Alter und den BMI („body mass index“) adjustiert.

Die Fall-/Kontroll-Gruppen umfassen 186 Fälle („atherogen“) und 185 Kontrollen („atheroprotektiv“) in Genf, sowie 82 Fälle und 75 Kontrollen in Berlin.

2.3 Definition und Bestimmung der Multi-Lokus-SNP-Genotypen

Es wurden 100 SNPs in 15 Kandidatenloci untersucht. Bei dieser Auswahl handelt es sich um bekannte Gene des Lipoproteinstoffwechsels, von denen angenommen wird, dass eine Funktionsveränderung der von diesen Genen kodierten Proteine zu einer Veränderung der Lipidkonzentrationen führt. Bei der Auswahl der SNPs lag der Schwerpunkt insbesondere auf der Auswahl von Polymorphismen in kodierenden und regulatorischen Bereichen der Gene, für die funktionelle Effekte in der Zelle vermutet wurden. Die Allele sollten eine Frequenz $> 3\%$ haben. Die Positionen der SNPs sollten den gesamten DNS-Abschnitt eines Gens überdecken. Die SNPs wurden aus der Literatur (Hubbard et al., 2002; Kerlavage et al., 2002; Pruitt & Maglott, 2001) und öffentlichen Datenbanken ausgewählt (Kapitel 3, Tabelle 3.2).

In der unabhängigen Stichprobe aus Genf wurden 275 SNPs aus 11 Kandidatengenen des Lipoproteinstoffwechsels ausgewählt (34 SNPs in 10 Genen waren identisch mit denen in Berlin). Anhand eines vorab-Screenings von 95 Individuen wurde die Frequenz untersuchter SNP-Positionen gemessen und häufige ($>3\%$) Polymorphismen in der gesamten Fall-Kontroll-Gruppe (371 Individuen mit Lipid-Konzentrationen im Randbereich) genotypisiert und ausgewertet.

2.4 Analysestrategie

- Die in den Labors gemessenen Plasmakonzentrationen der Cholesterinfraktionen: Gesamtcholesterin (TC), HDL-Cholesterin (HDL) und LDL-Cholesterin (LDL), Triglyzeride (TG) und das Verhältnis von LDL/HDL (LH-Ratio) werden als quantitative (reellwertige) phänotypische Merkmale betrachtet. Diese sind durch die genetische Konstitution der betreffenden Person, konkrete Lebensbedingungen und unvermeidliche zufällige Messfehler bestimmt. Es werden Logarithmen der Messwerte untersucht, die eine kompaktere Darstellung großer Messwertbereiche (Streuung) ermöglichen und die Annäherung an eine symmetrische Verteilung begünstigen. Diese Größen sind offensichtlich nicht unabhängig voneinander, denn sie werden z.T. auseinander berechnet (LH-Ratio aus dem Verhältnis von LDL zu HDL sowie LDL aus TC, HDL und TG

nach der Friedewald-Formel). Überdies sind sie im Stoffwechsel des Organismus auf komplexe Weise miteinander verbunden. Zur Vereinfachung werden die Phänotypen jedoch nicht als multivariate Größe betrachtet, sondern univariat modelliert.

- Das sichere physiologisch-biochemische Vorwissen über das Netzwerk des Lipoproteinstoffwechsels erlaubt die *a priori* Hypothese, dass die von unserer Arbeitsgruppe in ausführlichen Vorstudien ausgewählten Kandidatengene zu den wichtigsten Faktoren für dessen Funktion gehören. Die allelische Variation der zugehörigen Kandidatenorte wird als Faktor phänotypischer Variation untersuchter Merkmale des Lipoproteinstoffwechsels angesetzt.

- Die genetische Konstitution der Studienpopulation wird anhand von SNP-Genotypen und daraus abgeleiteten (geschätzten) SNP-Haplotypen beschrieben. Genkarten zeigen die Anordnung der SNP-Positionen in den Genen und die Markerdichte. Die Analyse des Kopplungsungleichgewichtes der Genorte wird auf die statistische Verteilung der Haplotypen in der Population zurückgeführt.

- Es wird angenommen, dass sich das betrachtete Merkmal als Funktion von genetischen Einflussgrößen und zufälligen Faktoren (Lebensbedingungen, Messbedingungen) darstellen lässt. Die Korrelation der Lipidwerte zwischen verwandten Personen wurde im Modell berücksichtigt. Diese Modellierung erfolgt auf dem Hintergrund von zufällig darstellbaren Einflussfaktoren.

- Die Modellierung basiert auf dem klassischen biometrisch-genetischen Ansatz, der besagt, dass in einer repräsentativen Population für die messbaren Faktoren in der Stichprobe ein mittlerer Einfluss auf das quantitative Merkmal angenommen wird (feste Effekte).

- Die Genotyp-Phänotyp-Beziehung wird simplifiziert als lineare Beziehung angenommen. Die Gendosis des selteneren Allels wurde additiv mit (0,1 oder 2) kodiert. Dies hat den Vorteil, dass die Stärke des Zusammenhangs mit einem Parameter (β) je Marker-Lokus modelliert werden kann. Diese Vereinfachung ist notwendig, um die Anzahl der Freiheitsgrade in dem Modell klein zu halten und damit eine größere Power für den statistischen Test zu erreichen.

- Sowohl Assoziation als auch Kopplung können nur dann mit diesem Ansatz nachgewiesen werden, wenn entweder der genetische Faktor das phänotypische Merkmal mitverursacht oder wenn er sich im Kopplungsungleichgewicht (LD) mit einem solchen Verursacher befindet. Das Ausmaß des nachzuweisenden Effektes hängt vom LD ab und lässt sich bei partiellem LD nicht vollständig darstellen.

- Untersuchte genetische Marker werden in multipler Regression linear modelliert. Dieser Ansatz kann bei Kollinearität zu verzerrten Schätzungen der Regressionsparameter führen. Innerhalb eines Gen-Lokus kann dieses Phänomen vernachlässigt werden, da nur die „gemeinsame“, globale Signifikanz analysiert wird. Wie sich Einzeleffekte untersuchter Marker innerhalb eines Lokus verteilen, wird nicht untersucht. Korrelationen zwischen Markern verschiedener Gene (größtenteils auf verschiedenen Chromosomen) treten nicht auf.

- Für alle genetischen Faktoren, die nicht explizit messbar sind, wird angenommen, dass ihre mittlere Wirkung eine positive Korrelation der phänotypischen Werte bei Individuen erzeugt, die miteinander verwandt sind, und dass die Stärke der Korrelation vom Verwandtschaftsgrad abhängt (zufällige Effekte).

- Im gemischten linearen Modell wird die intrafamiliäre Korrelation der Residuen entsprechend der Verwandtschaftsstruktur (je näher die Verwandtschaft, desto größer die Korrelation) untersucht. Der Einfluss so genannter Haushaltsfaktoren (in der Familie, ohne Berücksichtigung der Verwandtschaftsbeziehungen gleichmäßig wirksam) kann in speziell konstruierten Varianten des Modells getestet werden. Durch die Überlagerung familiärer genetischer Effekte und gemeinsamer familiärer Umwelteffekte ist es allerdings schwierig, die Varianzkomponente dieses Ansatzes entsprechend aufzuklären.

- Die Parameter der Funktion sollen plausibel geschätzt werden. Dazu wird die Maximum-Likelihood-Methode eingesetzt. Mit der ML-Methode wird ein Optimum der Wahrscheinlichkeitsdichte der Messwerte bedingt durch gegebene Parameter bestimmt. Für die Interpretation bedeutet dies, dass die Regressionskurve möglichst in der Mitte durch die Punktwolke der Messwerte führt („bester fit“).

- Zur Validierung werden Allelfrequenzen und LD in einer unabhängigen europäischen Studienpopulation untersucht. Die Multi-Lokus-Assoziationsanalyse wird im Rahmen eines Fall-/Kontroll-Designs geführt. Dazu wurde ein kombinierter qualitativer Phänotyp gebildet, der „atherogene“ bzw. „atheroprotektive“ Individuen anhand ihrer LDL- und HDL-Konzentrationen binär klassifiziert.

2.5 Globale Analyse des Kopplungsungleichgewichts

Kopplungsgleichgewicht (LE) bzw. Kopplungsungleichgewicht (LD) beschreibt den allelischen Zusammenhang polymorpher Positionen in der Population. Für q biallelische Marker gibt es 2^q mögliche Haplotypen, d.h. phasenspezifische allelische Kombinationen. Oft treten

weitaus weniger Haplotypen auf. Weicht die beobachtete Häufigkeit der Haplotypen in einer Population von der erwarteten Häufigkeit unter LE, d.h. dem Produkt der Allelfrequenzen einzelner Positionen ab, befinden sich die Positionen im Ungleichgewicht.

Die Analyse des LD zwischen Marker-Allelen basiert auf der Haplotypschätzung. Haplotypen wurden für jeden Gen-Lokus separat geschätzt. Zur Haplotypschätzung stehen verschiedene Algorithmen zur Verfügung. Dabei wird von q Genotypen $X = X_1, X_2, \ldots, X_q$, getrennt durch $q-1$ genomische Distanzen (Rekombinationswahrscheinlichkeiten) $\theta = \theta_{1,2}, \theta_{2,3}, \ldots, \theta_{q-1,q}$ ausgegangen.

Hier verwendete Haplotypschätzungen basieren auf Rohde & Fuerst (2001). Der Algorithmus von Rohde & Fuerst basiert auf einem EM-Algorithmus, der die Familienstruktur verwendet. Im Vergleich zu dem häufig benutzten Programm GENEHUNTER (Kruglyak & Lander, 1995), das den Lander-Green Algorithmus (Lander & Green, 1987) verwendet, werden dabei bessere Schätzungen bei mittlerer bis hoher Anzahl heterozygoter Positionen erzielt. Abhängig von der Größe des Gens und der Anzahl von SNPs pro Gen wurden Subhaplotypen gebildet.

Es gibt verschiedene Maße zur Charakterisierung des LD. Ein anerkanntes und häufig verwendetes Maß des LD zwischen zwei polymorphen Positionen ist der Korrelationskoeffizient nach Pearson r bzw. sein Quadrat r^2 (Δ^2) (Devlin & Risch, 1995). r^2 hat den Wert 1, wenn zwei SNPs auf demselben Chromosom liegen und nicht durch eine Rekombination unterbrochen sind. Der Wert ist kleiner als 1, wenn SNPs auf verschiedenen homologen Chromosomen auftreten oder falls eine ursprünglich enge Korrelation durch eine Rekombination zerstört worden ist. r bzw. r^2 sind komplexe Maße und unabhängig von der genomischen Distanz zwischen betrachteten Positionen. Dies ist sinnvoll, insofern die Stärke der Korrelation nicht notwendigerweise von der Distanz, sondern vom Zeitpunkt abhängt, an dem die Mutation entstanden ist (The International HapMap Consortium, 2005). Dies äußert sich in den Allelfrequenzen und Positionen der entsprechenden Mutationen in der Genealogie (Cordell & Clayton, 2005). Die Struktur paarweisen LDs wurde hier lokusweise mit r berechnet. Dazu werden aus den SNP-Genotypen zweier biallelischer Positionen A und B (AA/Aa/aa und BB/Bb/bb) zunächst Haplotypen (AB/Ab/aB/ab) geschätzt. r kann aus der Vierfeldertafel mit $r = (p_{AB}p_{Ab} - p_{aB}p_{ab})/\sqrt{p_{A.}p_{a.}p_{.B}p_{.b}}$ berechnet werden, wobei p_{xy} die Zellhäufigkeiten und $p_{x.}$ die Randhäufigkeiten darstellen. Die Signifikanz des paarweisen Zusammenhangs der phasenspezifischen SNP-Allele wurde in einem χ^2-Unabhängigkeits-Test oder, bei kleinen Zellhäufigkeiten, mit Fishers Exaktem Test bestimmt. Dabei werden die Häufigkeiten der Allele in der Vierfeldertafel miteinander verglichen. Das Ergebnis ist eine symmetrische Matrix der Korrelationen und Signifikanzen aller paarweisen SNP-Vergleiche separater Loci. Die Werte wurden in Form einer oberen Dreiecksmatrix visualisiert.

Der Vergleich der multiplen LD-Struktur in den zwei unabhängigen Stichproben (Genf;

Berlin) wurde auf der Basis eines Determinantenkriteriums geführt. Die Determinante der zweidimensionalen Korrelationsmatrix ist $(1 - r^2)$. Sie ist ≈ 1 im Fall von Kopplungsgleichgewicht (zufällige Kombination aller möglichen Haplotypen) und ≈ 0 im Fall von LD (Abweichung von der zufälligen Paarung). Das Analogon im mehrdimensionalen Fall ist die Determinante der mehrdimensionalen Korrelationsmatrix R. LD von SNPs eines Gen-Lokus wurde gegen die Nullhypothese (volles LE, d.h. Einheitsmatrix) getestet. Morrison (2005) beschreibt eine passende Teststatistik mit $-log(Det(R))$. Die empirische Verteilung der Teststatistik unter der Nullhypothese wurde durch 1000-fache Permutation realisiert. Dazu wurden die einzelnen zu einem Gen-Lokus gehörenden SNP-Vektoren unabhängig voneinander permutiert. Die Abweichung der LD-Struktur vom Gleichgewicht wurde auf der Basis des Wertes $-log(det(R))/q$ charakterisiert, der die Anzahl von SNPs (q) berücksichtigt. Werte < 0.1 deuten auf schwaches LD, Werte zwischen 0.1 und 0.4 auf mittleres LD und Werte zwischen 0.4 und 1.0 auf starkes LD zwischen untersuchten genetischen Markern. (Die Skala ist heuristisch angesetzt.)

Die Gleichheit der Kovarianzstruktur zwischen Genf und Berlin S_i ($i = 1, 2$ für Genf, Berlin) basierend auf gemeinsamen SNPs pro Gen-Lokus, für jeweils N_i SNP-Genotypen, wurde nach der folgenden Teststatistik (Morrison, 2005) analysiert:

$$M = (n_1 + n_2)ln(Det(S)) - n_1 ln(Det(S_1)) - n_2 ln(Det(S_2)), \tag{2.1}$$

wobei $n_1 = N_1 - 1$, $n_2 = N_2 - 1$. $S = n_1 S_1 + n_2 S_2)/(n_1 + n_2)$ das gewichtete Mittel bezeichnet. Die empirische Verteilung von M unter der Nullhypothese wurde durch 1000-fache Permutation des gemessenen SNP-Vektors realisiert.

2.6 Nachweis phänotypischer Assoziation in Familien unter Anwendung des biometrisch-genetischen Modells

Das hier verwendete statistische Modell ist von dem klassischen biometrisch-genetischen Modell polygener Einflüsse (dargestellt in Falconer & Mackay (1996); Lange (1997)) unter vereinfachten Annahmen abgeleitet. Dieser Ansatz geht zurück auf Fisher (1918), der die Biometrie mit den Mendelschen Gesetzen der Vererbung kombinierte.

Fisher definiert den quantitativen Phänotyp (P) als Funktion genotypischer Effekte (G) und der umweltbezogenen Abweichung (E):

$$P = G + E. \tag{2.2}$$

Die umweltbezogene Abweichung ist im Mittel = 0, so dass der phänotypische Wert im Mittel gleich dem genotypischen Wert ist. Das Populationsmittel bezieht sich demnach gleichermaßen auf P bzw. G. Dabei wird angenommen, dass die Umwelteinflüsse über die Generationen hinweg konstant bleiben, so dass sich das Populationsmittel bei gleicher genetischer Konstitution nicht ändert.

Fisher erweitert das Modell 2.2 im Rahmen einer Varianzanalyse, wobei die phänotypische Gesamtvarianz in einzelne Komponenten aufgeteilt werden kann.

$$Var(P) = Var(G) + Var(E) + 2Cov(GE) \tag{2.3}$$

wobei $Var(P)$ die phänotypische Varianz (σ^2), $Var(G)$ die genotypische Varianz (σ^2_{poly}) und $Var(E)$ die Umweltvarianz (σ^2_{env}), bedingt durch den Einfluss nicht-genetischer Faktoren und Messfehler bezeichnen. Es wird angenommen, dass genotypische und umweltbezogene Abweichungen unkorreliert sind, d.h. dass die Kovarianz zwischen G und E ($Cov(GE)$) = 0. Im Weiteren werden die Varianzkomponenten als theoretische Parameter der Modelle mit σ^2 und als Schätzer mit v bezeichnet.

Das biometrisch-genetische Modell beschreibt die genetische Architektur eines quantitativen Merkmals. Angenommen, je zwei Allele segregieren an jedem Marker-Lokus, der zur phänotypischen Variation beiträgt, dann sind die individuellen Genotypen durch homozygote oder heterozygote Effekte jedes Lokus und Interaktionen zwischen den Loci spezifiziert. Es wird angenommen, dass „meiotic drift“ keine Rolle spielt, d.h. dass heterozygote Eltern eines ihrer beiden Allele mit jeweils gleicher Wahrscheinlichkeit an die Nachkommen weitergeben. Weiterhin wird „random mating“, d.h. Hardy-Weinberg-Gleichgewicht, untersuchter Marker vorausgesetzt.

Aussagen zu der genetischen Struktur des quantitativen Merkmals können durch Zerlegung der phänotypischen Gesamtvarianz in Komponenten additiver, dominanter, epistatisch genetischer Varianz, Varianz der Genotyp-Umwelt-Interaktionen und zusätzlicher Umweltvarianz abgeleitet werden. Das Verhältnis genetisch bestimmter Varianz an phänotypischer Gesamtvarianz wird auch als „Heritabilität“ bezeichnet. Die Heritabilität ergibt sich als Verhältnis

$$\text{Heritabilität} = \frac{\sigma^2_{poly}}{\sigma^2}. \tag{2.4}$$

Die Varianzkomponentenschätzungen sind abhängig von der untersuchten Studienpopulation, (1) da die genetischen Parameter des Modells von den Allelfrequenzen beteiligter Marker-Loci abhängen und (2) bedingt durch umweltbezogene Unterschiede zwischen Populationen (Mackay, 2001). Im Fall von Populationsstratifikation kann sich die Verteilung der Genotypen (und Phänotypen) zwischen Sub-Populationen unterscheiden und zu „falschen“ Assoziationsergebnissen führen.

Das Standardmodell polygener Einflüsse geht von einem quantitativen Merkmal aus, das durch eine große Anzahl genetischer Einflussparameter bestimmt wird, die unabhängig und additiv wirken. Ausgehend von einer Familie i mit J_i Mitgliedern bezeichne g_j^h, $j, \ldots, J_i$ den Beitrag des Marker-Lokus h, $h = 1, \ldots, q$ SNPs oder SNP-Haplotypen, zum phänotypischen Wert des Individuums j. Der genotypische Wert $g_j = \sum_{h=1}^{q} g_j^h$ der Person j ist Teil des Vektors $G = (g_1, \ldots, g_{J_i})^T$ der Werte für die Familie. Wenn der additive Einfluss der h verschiedenen Marker-Loci g^h auf den Phänotyp von gleicher Größenordnung ist, dann nähert sich die Summe G der Einflüsse nach den Grenzwertsätzen der Wahrscheinlichkeitstheorie der multivariaten Normalverteilung. Weiterhin impliziert die Unabhängigkeit der verschiedenen Marker-Loci voneinander $Cov(g_j, g_k) = \sum_{h=1}^{q} Cov(g_j^h, g_k^h)$ mit $j, k = 1, \ldots, J_i$ (Lange, 1997).

Die Kovarianz zwischen Verwandten kann in Werten und Frequenzen der Genotypen ausgedrückt werden. Die mittlere Wahrscheinlichkeit, zwei Allele IBD („identical by descent“ = identisch geerbt) zu haben, wird durch den theoretischen Kinship-Koeffizienten ϕ_{jk} zwischen jedem Verwandtschaftspaar j und k ausgedrückt. Der theoretische Kinship-Koeffizient ist ein allgemeines Verwandtschaftsmaß und wird innerhalb eines Stammbaumes rekursiv berechnet (Lange, 1997). Dieser Koeffizient ϕ_{jk} beträgt für zwei Individuen j und k: 0, wenn sie nicht miteinander verwandt sind, 1/4 zwischen Vollgeschwistern sowie für ein Mutter(Vater)-Kind-Paar, 1/8 zwischen Halbgeschwistern sowie für Onkel(Tante)-Neffe(Nichte)-Paar, 1/16 bzw. 1/64 zwischen Cousins 1.Grades bzw. 2.Grades. Abgeleitet daraus hat die Kovarianzmatrix einer Kernfamilie, beispielsweise bestehend aus Eltern und zwei Kindern, die folgende Struktur:

$$Cov(g_j, g_k) = 2 \begin{pmatrix} \text{Mutter} & 1/2 & 0 & 1/4 & 1/4 \\ \text{Vater} & 0 & 1/2 & 1/4 & 1/4 \\ \text{Tochter} & 1/4 & 1/4 & 1/2 & 1/4 \\ \text{Sohn} & 1/4 & 1/4 & 1/4 & 1/2 \end{pmatrix} \sigma^2_{poly}. \tag{2.5}$$

σ^2_{poly} bezeichnet die über alle beteiligten Loci summierte genetische Varianz. Es wird angenommen, dass σ^2_{poly} für alle Familien den gleichen Wert hat. Die Matrix der theoretischen Kinship-Koeffizienten ϕ_{ij} ist symmetrisch und positiv definit. Allgemein ergibt sich die Kovarianz zwischen einem beliebigen Verwandschaftspaar als

$$Cov(g_j, g_k) = 2\phi_{jk}\sigma^2_{poly}. \tag{2.6}$$

Unter Einbeziehung von Umwelteffekten lassen sich die gemessenen Lipidphänotypen als Summe $y_j = g_j + e_j$ des genetischen g_j und des umweltbedingten Beitrages e_j darstellen. Mit der Annahme, dass der Zufallsvektor $E = (e_1, \ldots, e_{J_i})^T$ unabhängig von G multivariat

normal verteilt ist mit dem Erwartungswert-Vektor μ und der Varianz Λ, dann ist $Y = (y_1, \ldots, y_{J_i})^T$ ebenfalls normalverteilt mit $E(Y) = \mu$ und $Var(Y) = 2\Phi\sigma^2_{poly} + \Lambda$. Λ wird hier in der einfachsten Form mit $\Lambda = \sigma^2_{env} I$ angenommen, wobei I die $J_i \times J_i$-Einheitsmatrix bezeichnet. Die Varianz-Kovarianzmatrix für eine Familie i kann mit

$$\Omega_i = \begin{cases} \sigma^2_{poly} + \sigma^2_{env} & \text{falls } j = k, \\ 2\phi_{ijk}\sigma^2_{poly} & \text{falls } j \neq k. \end{cases} \tag{2.7}$$

angegeben werden, wobei $j, k = 1, \ldots, J_i$.

In einer zufälligen Stichprobe von M Familien sind die Beobachtungen unverwandter Familien unabhängig. Der Vektor y_i der gemessenen quantitativen Phänotypen innerhalb einer Familie hat eine Korrelationsstruktur ausgedrückt durch die $J_i \times J_i$ Matrix Ω_i. Die in dieser Arbeit untersuchten Familien haben unterschiedliche Größe und Verwandtschaftsstruktur, d.h. es liegt ein unbalanciertes Design vor. Die Matrix Ω der gesamten Stichprobe mit N Individuen hat die Form einer Blockdiagonalmatrix, bestehend aus den familienspezifischen Matrizen Ω_i, $(i = 1, \ldots, M)$.

Mit diesem allgemeinen Ansatz werden die Messdaten an ein Modell angepasst, in dem gleichzeitig Einflüsse auf den Mittelwert (μ_0,β), in der Terminologie gemischter Modelle als „feste Effekte“ bezeichnet, und Varianzkomponenten (σ^2_{poly},σ^2_{env}) als „zufällige Effekte“ geschätzt werden.

Die Likelihood-Funktion ist die Wahrscheinlichkeitsdichte für die Daten, gegeben die Parameter. Sie wird jedoch als eine Funktion der Parameter mit festen Daten anstelle einer Funktion der Daten mit festen Parametern betrachtet, d.h.

$$L(\mu, \Omega|Y) = p(Y|\mu, \Omega), \tag{2.8}$$

wobei L die Likelihood, p die Wahrscheinlichkeitsdichte und Y den N-dimensionalen Phänotypvektor, $N = \sum_{i=1}^{M} J_i$, bezeichnen. Mit der Annahme von Unabhängigkeit der zufälligen Effekte g_i, $i = 1, \ldots, M$ und e_{ij}, $j = 1, \ldots, J_i$ ergibt sich

$$L(\mu, \Omega|Y) = \prod_{i=1}^{M} p(y_i|\mu, \Omega). \tag{2.9}$$

(Pinheiro & Bates, 2000).

Das Produkt unabhängig normalverteilter Größen führt auf eine mehrdimensionale Normalverteilung und die Log-Likelihood-Funktion des Modells hat die Form

$$lnL(\mu, \Omega|Y) = -\frac{1}{2} \sum_{i}^{M} \left[ln|\Omega_i| + (y_i - \mu_i)' \Omega_i^{-1} (y_i - \mu_i) \right]. \tag{2.10}$$

Nach dem Maximum-Likelihood-Prinzip sucht man nun denjenigen Parametersatz, der den gemessenen Daten die höchste Wahrscheinlichkeit verleiht. Die Vorhersage des Phänotyps nach der ML-Schätzung der Parameter liefert die „plausibelste Anpassung“ der Daten an das Modell. Je höher der ML-Wert, desto höher ist die Anpassungsgüte.

Auf dieser Basis werden Likelihood-Ratio-Tests der Form $2(lnL_A - lnL_0)$ für verschiedene Varianten des Modells konstruiert. L_0 bezeichnet die Likelihood-Funktion unter der Nullhypothese H_0, wobei untersuchte Parameter festgesetzt (beschränkt) werden. (Man nimmt z.B. an, dass die Parameterwerte = 0, die Einflussgrößen also nicht assoziiert sind.) L_A bezeichnet die Likelihood-Funktion unter der Alternativhypothese H_A, wobei untersuchte Parameter frei geschätzt werden.

In großen Stichproben ist $2ln(L_A/L_0)$ asymptotisch χ^2-verteilt mit der Anzahl von Freiheitsgraden äquivalent zur Anzahl getesteter Parameter. Dies gilt allgemein in der Modellierung quantitativer Phänotypen anhand von Geschwistern oder Großfamilien (Fulker et al., 1999). Der Likelihood-Ratio-Test ist relativ robust gegenüber Abweichungen von der Normalverteilung, verursacht u.a. durch Hauptgeneffekte, Wechselwirkung genetischer und umweltbezogener Effekte oder durch Selektionsfehler (Allison et al., 1999). Durch Permutation der Phänotypwerte wurde die Nullhypothese simuliert und damit die Signifikanz geschätzter Modelle zusätzlich abgesichert.

Assoziation bedeutet im hier betrachteten Zusammenhang, dass ein quantitativ messbares Merkmal von der Ausprägung eines genetischen Merkmals (oder Merkmalskomplexes) abhängt. Der Phänotyp ist die abhängige Variable und der Vektor mit den Werten der betrachteten Genotypen (SNPs, Haplotypen) die unabhängige Variable.

Der Test auf Multi-Lokus-Assoziation basiert auf dem Vergleich von Modellen mit und ohne Einfluss multipler genotypischer Information an multiplen Marker-Loci. Unter der Nullhypothese H_0 (keine Assoziation) besteht das Modell der festen Effekte nur aus dem Populationsmittel μ_0 des quantitativen Merkmals, polygene Einflüsse auf den Phänotyp werden in die zufällige Komponente polygener Restvarianz modelliert. Unter der Alternativhypothese H_A werden alle gemessenen genetischen Effekte linear in den Mittelwert modelliert. Die asymptotische Parameterschätzung für den Mittelwert ergibt sich als

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu = \mu_0 + X\beta. \end{aligned} \tag{2.11}$$

X bezeichnet die konstante $N \times q$ - Matrix der q Genotypen (SNPs oder Haplotypen) in N Individuen. SNP-Genotypen werden dabei nach dem Vorhandensein auf den zwei Chromosomen mit 0/1/2 (Alleldosis) der Individuen bewertet. Haplotypen an einem Lokus können äquivalent zu den SNPs ebenfalls entsprechend der Häufigkeit ihrer Haplotyp-Allele

kodiert werden. Somit erhält jedes Individuum den Wert 0/1/2, der angibt, wie oft der entsprechende Haplotyp individuell vorhanden ist. Die Haplotypanalyse erlaubt den Test kombinierter Effekte multipler Sequenz-Varianten eines Chromosoms auf die Variation des untersuchten Merkmals. Marker mit einer Allelfrequenz $< 5\%$ wurden ausgeschlossen und Marker im paarweisen LD ($|r| >= 0.7$) gruppiert.

Es handelt sich um ein lineares Modell ohne Dominanzeffekte. β ist ein $q \times 1$ - Vektor der festen Parameter und beschreibt den additiven Effekt eines genotypischen Merkmals. Die zufälligen Effekte werden in Form einer Blockdiagonalmatrix der familienspezifischen Matrizen Ω_i (2.7) modelliert. Unter H_0 liefert die Schätzung der Varianzkomponente v_{poly} ein Maß der Heritabilität des untersuchten phänotypischen Merkmals.

Neben dem Likelihood-Ratio-Test, in dem die Likelihooddifferenz zwischen den Modellen mit und ohne Einschluss multipler genetischer Marker in einem χ^2-Test mit der theoretischen Verteilung verglichen wird, wurde die Signifikanz des Assoziationsmodells in einem Permutationstest geprüft. Dazu wurde zur Realisierung der Nullhypothese der Phänotyp y permutiert und q multiple (nicht assoziierte) Genotypen modelliert. Das empirische Signifikanzniveau ergibt sich aus dem Vergleich der Likelihooddifferenz - zwischen den Modellen ohne und mit Einschluss multipler SNPs/SNP-Haplotypen - des permutierten Modells und des originalen Modells in 1000 Simulationen. Damit ist der Signifikanzwert des Assoziationsmodells gegen Abweichungen von der mehrdimensionalen Normalverteilung abgesichert (Voraussetzung für den χ^2-Test).

Die Variabilität der Varianzkomponentenschätzungen wurde in Bootstrap-Simulationen verifiziert. Bootstrap-Verfahren werden angewendet, um die Verteilung der Parameter eines Modells zu approximieren. Um die Familienstruktur in den Bootstrap-Stichproben widerzuspiegeln, wurde das Verfahren des Block-Bootstrap angewendet (Efron & Tibshirani, 1998). Dabei werden mit Zurücklegen Blöcke (hier Familien) aus den Orginaldaten gezogen. Mit unterschiedlichen Familiengrößen resultieren entsprechend ungleiche Dimensionen der Bootstrap-Stichproben. Efron & Tibshirani führen für diesen Fall einen Korrekturfaktor ein. Basierend auf den B Bootstrap-Stichproben wurde die empirische Verteilung der Varianzkomponentenschätzungen ermittelt und das 95%-Konfidenzintervall für v_{poly} und v_{env} angegeben.

Die Bedeutung einzelner Gen-Loci innerhalb des Gesamtmodells (2.11) wird partiell getestet. Ein Gen-Lokus wird im Multi-Lokus-Modell durch s multiple Genotypen als Regressionsparameter repräsentiert. Die Hypothesen dieses Ansatzes ergeben sich als

$$\begin{aligned} H_0 &: \mu = \mu_0 + X^{q-s}\beta^{q-s}, \\ H_A &: \mu = \mu_0 + X^{q}\beta^{q} \end{aligned} \tag{2.12}$$

wobei unter H_0 $(q - s)$ gemessene Multi-Lokus-Genotypen modelliert werden. X^{q-s} bezeichnet die Matrix der $(q - s)$ SNPs. Alternativ wird der genotypische Einfluss aller q Genotypen (X^q) betrachtet.

Ein Schätzer des Beitrags eines Gen-Lokus an genetisch bedingter phänotypischer Varianz ist die Differenz der Komponenten polygener Restvarianz bei Einschluss/Ausschluss gemessener Lokus-spezifischer genetischer Marker als feste Effekte (Boerwinkle & Sing, 1986; Wijsman & Nur, 2001). Hier wird der partielle Anteil σ^2_{locus} eines Gen-Lokus aus der Differenz der Komponenten polygener Restvarianz der Modelle unter H_0 und H_A (2.12) als

$$\sigma^2_{locus} = \sigma^2_{poly}(\text{Modell unter} H_0) - \sigma^2_{poly}(\text{Modell unter} H_A) \tag{2.13}$$

geschätzt.

Populationsstratifikation wurde auf der Basis untersuchter genetischer Marker im Vorfeld der multiplen Tests geprüft. Abecasis et al. (2000b) haben im Rahmen des beschriebenen gemischten Modells einen Assoziationstest für Großfamilien vorgeschlagen, der robust gegen Populationsstratifikation ist. Dabei wird die Gesamtassoziation in den Effekt innerhalb der Familien (ausgedrückt durch den individuellen Abstand vom genotypischen Familienmittel) und zwischen den Familien (Abstand der Familienmittel vom Gesamtmittel) aufgespalten. Hinweise auf Stratifikation bestehen, wenn der Assoziationseffekt nur zwischen den Familien nachweisbar ist. Der Ansatz von Abecasis et al. (2000b) stellt eine Erweiterung der Ansätze von Fulker et al. (1999) und Sham et al. (2000) dar, die Geschwisterpaare bzw. mehrere Geschwister betrachtet haben.

Zur Modellierung des oben beschriebenen gemischten Modells mit Kinship-Struktur wurde zunächst das Programm QTDT (Abecasis et al., 2000a) in modifizierter Anwendung und später das R-Paket „Kinship“ (Atkinson & Therneau, 2004) verwendet.

2.7 Vergleich der genetischen Faktoren in zwei ethnisch verschiedenen Populationen

Zur Validierung der Assoziationsbefunde aus der familienbasierten Analyse wurde eine Vergleichsstudie in einer unabhängigen Stichprobe aus der Schweiz durchgeführt. Gen-basiert soll getestet werden, ob untersuchte Kandidatengene gleichermaßen auf den Lipidphänotypen wirken.

Aufgrund des unterschiedlichen Studiendesigns beider Stichproben wurde der Vergleich auf der Basis unabhängiger (unverwandter) Individuen geführt. Die Genfer Studie beruht auf dem Vergleich von Fällen und Kontrollen klassifiziert nach einem qualitativen dichotomisierten Phänotyp, der das atherogene Risiko eines Individuums beschreibt. Deutsche

Individuen wurden entsprechend klassifiziert.

Als genetische Marker wurden SNP-Genotypen betrachtet und nach Abwesenheit (=0) oder Anwesenheit (=1) des seltenen Alleles kodiert. Diese Kodierung basiert auf der Originalstudie von Morabia et al. (2003b) und unterstellt einen dominanten allelischen Effekt.

Odds Ratios (OR) und dazugehörige 95%-Konfidenzintervalle (KI) einzelner SNPs, adjustiert für den BMI und das Alter der Probanden, wurden in logistischer Regression mit dem qualitativen Lipid-Phänotypen als abhängiger Variable ermittelt. Dabei wurden Alter und BMI in den Genfer Modellen als lineare Kovariablen modelliert und das Geschlecht als SNP-Geschlechts-Interaktionsterm einbezogen. Im Fall eines nominal signifikanten Interaktionseffektes ($p < 0.10$) wurden die Modelle stratifiziert nach dem Geschlecht, sonst nicht.

Die Assoziationseffekte identischer SNPs in den unabhängigen Stichproben aus Genf und Berlin wurden auf der Basis des Zweistichproben-t-Tests miteinander verglichen.

Der Gen-Lokus-Effekt wurde für jedes Gen separat in einem multiplen Assoziationsansatz im Rahmen einer multiplen logistischen Regression umgesetzt. Die SNP-Genotypen gehen als unabhängige Variablen ein. Sie werden linear modelliert und Interaktionsterme (SNP*Geschlecht) eingeschlossen.

Die große Anzahl potentiell assoziierter SNP-Varianten (244 in Genf und 74 in Berlin) gegenüber dem nicht deutlich größeren Stichprobenumfang der Fall-Kontroll-Gruppen (371 in Genf, 157 in Berlin) erforderte für eine asymptotische Parameterschätzung eine Reduktion der Einflussgrößen. Durch schrittweise Variablenselektion wurde ein Sub-Modell nach Akaikes Informationskriterium (AIC) bestimmt. $AIC = -2lnL + 2q$, wobei lnL die maximierte Log-Likelihood der Daten nach Anpassung des Modells und q die Anzahl der Kovariablen plus Interzept bezeichnen. Das Ergebnis der Screening-Prozedur war ein reduziertes SNP und SNP/Interaktions-Set für jeden Gen-Lokus. Die Signifikanz des multiplen logistischen Regressionsmodells für einen Lokus ergibt sich als p-Wert eines Likelihood-Ratio-Tests zwischen dem Multi-Lokus-Modell (Einschluss selektierter Parameter) und dem „Null Modell“ (nur Interzept).

Die Güte der Anpassung wurde mit R^2, d.h. dem „kumulativen adjustierten Bestimmtheitsmaß“ nach Nagelkerke (1991) beschrieben. Dieser Wert variiert zwischen 0 und 100% und beschreibt den Anteil der Variation, der durch das Modell erklärt werden konnte. Ein unverzerrter Schätzer für R^2 wurde anhand von Bootstrap-Simulationen berechnet (Harrell, 1999).

Der prädiktive Wert angegebener Modelle wurde außerdem in 10-facher Kreuzvalidierung (Efron & Tibshirani, 1998) anhand des Anteils korrekt vorhergesagter Phänotypen (Wahrscheinlichkeit $p > 0.5$ für Fälle und $p > 0.5$ für Kontrollen) bestimmt. Die Signifikanz wurde theoretisch auf der Basis der Bernoulliverteilung angegeben. Wegen der großen

umweltbedingten und genetischen Hintergrundvarianz liegt der Anteil korrekter Prädiktionen für einen einzelnen Lokus zwischen 50% und schätzungsweise 65%. Solche Prozentzahlen sind leicht interpretierbar, ihre Schätzung jedoch möglicherweise numerisch instabil und abhängig von der gewählten Klassifikationsschranke. Eine weniger anfällige Maßzahl prädiktiver Genauigkeit ist der Wert D_{xy} nach Somers (1962). D_{xy} misst die Korrelation zwischen der beobachteten binären Zielgröße und den vorhergesagten Wahrscheinlichkeiten (Harrell, 1999).